Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/JP2005/017696

International filing date: 27 September 2005 (27.09.2005)

Document type: Certified copy of priority document

Document details: Country/Office: JP
Number: 2005-131992
Filing date: 28 April 2005 (28.04.2005)

Date of receipt at the International Bureau: 17 November 2005 (17.11.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 Date of Application:	２００５年　４月２８日
出 願 番 号 Application Number:	特願２００５－１３１９９２
パリ条約による外国への出願に用いる優先権の主張の基礎となる出願の国コードと出願番号 The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is	ＪＰ２００５－１３１９９２
出 願 人 Applicant(s):	松下電器産業株式会社

２００５年１１月　２日

特許庁長官
Commissioner,
Japan Patent Office

中嶋 誠

【書類名】　特許願
【整理番号】　7048070042
【提出日】　平成17年 4月28日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/30
【発明者】
　【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内
　【氏名】　稲葉　光昭
【発明者】
　【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内
　【氏名】　菅野　祐司
【特許出願人】
　【識別番号】　000005821
　【氏名又は名称】　松下電器産業株式会社
【代理人】
　【識別番号】　100097445
　【弁理士】
　【氏名又は名称】　岩橋　文雄
【選任した代理人】
　【識別番号】　100103355
　【弁理士】
　【氏名又は名称】　坂口　智康
【選任した代理人】
　【識別番号】　100109667
　【弁理士】
　【氏名又は名称】　内藤　浩樹
【先の出願に基づく優先権主張】
　【出願番号】　特願2004-345392
　【出願日】　平成16年11月30日
【手数料の表示】
　【予納台帳番号】　011305
　【納付金額】　16,000円
【提出物件の目録】
　【物件名】　特許請求の範囲　1
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9809938

【書類名】特許請求の範囲
【請求項１】
構造化文書を管理するデータベース構築装置において、
構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行う入力文書解析部と、
前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各要素名に対してユニークな要素名ＩＤを割り当てて要素名辞書に登録する要素名登録部と、
前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて祖先パス名辞書に登録する祖先パス名登録部と、
前記入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして要素出現情報格納部に登録し、かつ、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして祖先パス出現情報格納部に登録する出現情報登録部と、
を有することを特徴とするデータベース構築装置。
【請求項２】
前記入力文書解析部の解析結果に基づいて、構造化文書に出現する各属性名に対してユニークな属性名ＩＤを割り当てて属性名辞書に登録する属性名登録部を有し、
前記出現情報登録部が、前記入力文書解析部の解析結果に基づいて、着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順の情報を少なくとも含む属性出現情報を、属性名ＩＤをキーとして属性出現情報格納部に登録することを特徴とする請求項１に記載のデータベース構築装置。
【請求項３】
前記出現情報登録部が、前記入力文書解析部の解析結果に基づいて、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと属性名ＩＤと分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納部に登録することを特徴とする請求項１に記載のデータベース構築装置。
【請求項４】
前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順と空要素順の情報を少なくとも含み、
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順と空要素順の情報を少なくとも含むことを特徴とする請求項１に記載のデータベース構築装置。
【請求項５】
前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順と空要素順の情報を少なくとも含み、
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順と空要素順の情報を少なくとも含み、
前記属性出現情報は、着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順と空要素順の情報を少なくとも含むことを特徴とする請求項２に記載のデータベース構築装置。
【請求項６】
前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順と空要素順の情報を少なくとも含み、
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順と空要素順の情報を少なくとも含み、
前記属性出現情報は、着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順と空要素順の情報を少なくとも含み、

前記テキスト出現情報は、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと属性名ＩＤと分岐順と空要素順の情報を少なくとも含むことを特徴とする請求項３に記載のデータベース構築装置。
【請求項７】
前記祖先パス名登録部は、前記構造化文書に出現する各祖先パス名を１つ以上に分割した各々の部分祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて祖先パス名辞書に登録することを特徴とする請求項１に記載のデータベース構築装置。
【請求項８】
前記要素出現情報格納部に同じ要素名ＩＤをキーにして登録されている前記要素出現情報のエントリ群と、前記祖先パス出現情報格納部に同じ祖先パス名ＩＤをキーにして登録されている前記祖先パス出現情報のエントリ群とに対して、文書番号と文字位置以外の１つ以上の情報項目の値が共通するエントリ同士をグループ化する出現情報グループ化部を有することを特徴とする請求項１に記載のデータベース構築装置。
【請求項９】
構造化文書を管理するデータベース検索装置において、
構造化文書に出現する各要素名に対してユニークな要素名ＩＤを登録した要素名辞書と、
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを登録した祖先パス名辞書と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして格納した要素出現情報格納部と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして格納した、祖先パス出現情報格納部と、
検索式を入力するための検索条件入力部と、
前記要素名辞書と前記祖先パス名辞書とを参照して、前記入力された検索式を内部条件式に変換する検索条件解析部と、
前記検索条件解析部の出力した内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報および、前記祖先パス出現情報格納部からの祖先パス出現情報から検索結果群を求める出現情報取得部と、
を有することを特徴とするデータベース検索装置。
【請求項１０】
属性名ＩＤと対応する属性名の記録された属性名辞書と、
着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順の情報を少なくとも含む属性出現情報を、属性名ＩＤをキーとして格納した属性出現情報格納部とを有し、
前記検索条件解析部が、前記要素名辞書と前記祖先パス名辞書と前記属性名辞書とを参照して、前記検索条件入力部から入力された検索式を内部条件式に変換し、前記出現情報取得部が、前記検索条件解析部の出力した内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報、前記祖先パス出現情報格納部からの祖先パス出現情報および、前記属性出現情報格納部からの属性出現情報から検索結果群を求めることを特徴とする請求項９に記載のデータベース検索装置。
【請求項１１】
要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと属性名ＩＤと分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとして格納した、テキスト出現情報格納部とを有し、
前記出現情報取得部が、前記検索条件解析部の出力した内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報、前記祖先パス出現情報格納部からの祖先パス出現

情報、前記属性出現情報格納部からの属性出現情報および、前記テキスト出現情報格納部からのテキスト出現情報から検索結果群を求めることを特徴とする請求項９に記載のデータベース検索装置。
【請求項１２】
前記出現情報取得部は、前記要素出現情報格納部における指定要素名ＩＤのエントリ数と、前記祖先パス出現情報格納部における指定祖先パス名ＩＤのエントリ数の大小を比較し、少ない方の出現情報を参照するようにして検索結果群を求めることを特徴とする請求項９乃至１１のいずれかに記載のデータベース検索装置。
【請求項１３】
構造化文書を管理するデータベース構築方法において、
構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行うステップと、
前記解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名ＩＤを割り当てて要素名辞書に登録するステップと、
前記解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて祖先パス名辞書に登録するステップと、
前記解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして要素出現情報格納部に、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして祖先パス出現情報格納部にそれぞれ登録するステップと、を有することを特徴とするデータベース構築方法。
【請求項１４】
前記要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順と空要素順の情報を少なくとも含み、
前記祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順と空要素順の情報を少なくとも含むことを特徴とする、請求項１３に記載のデータベース構築方法。
【請求項１５】
前記祖先パス名辞書に登録するステップは、構造化文書に出現する各祖先パス名を１つ以上に分割した各々の部分祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて登録するステップであり、
前記要素出現情報には、単一の祖先パス名ＩＤの代わりに１つ以上の祖先パス名ＩＤの列を含み、
前記祖先パス出現情報格納部には、単一の祖先パス名ＩＤの代わりに１つ以上の祖先パス名ＩＤの列をキーとして前記祖先パス出現情報を登録することを特徴とする、請求項１３記載のデータベース構築方法。
【請求項１６】
前記要素出現情報格納部に同一の要素名ＩＤをキーとして登録され、文書番号と文字位置以外の情報項目の値が共通であるような前記要素出現情報のエントリ同士をグループ化し、前記祖先パス出現情報格納部に同一の祖先パス名ＩＤをキーとして登録され、文書番号と文字位置以外の情報項目の値が共通であるような前記祖先パス出現情報のエントリ同士をグループ化するステップを有することを特徴とする請求項１３記載のデータベース構築方法。
【請求項１７】
構造化文書を管理するデータベース検索方法において、
構造化文書に出現する各要素名に対してユニークな要素名ＩＤを登録した要素名辞書と、
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを登録した祖先パス名辞書と、
前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして格納した要素出現情報格納部と、

前記構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして格納した、祖先パス出現情報格納部と、
検索式を入力するためのステップと、
前記要素名辞書と前記祖先パス名辞書とを参照して、前記入力された検索式を内部条件式に変換するステップと、
前記内部条件式にしたがって、前記要素出現情報格納部からの要素出現情報および、前記祖先パス出現情報格納部からの祖先パス出現情報から検索結果群を求めるステップと、
を有することを特徴とするデータベース検索方法。
【請求項１８】
構造化文書を管理するデータベース装置において、
構造化文書に出現する各要素名に対してユニークな要素名ＩＤを記憶する要素名辞書と、
前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを記憶する祖先パス名辞書と、
構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行う入力文書解析部と、
前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各要素名に対してユニークな要素名ＩＤを割り当てて前記要素名辞書に登録する要素名登録部と、
前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて前記祖先パス名辞書に登録する祖先パス名登録部と、
文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして記憶する要素出現情報格納部と、
文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして記憶する祖先パス出現情報格納部と、
前記入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、前記着目要素の要素名ＩＤをキーとして前記要素出現情報格納部に登録し、かつ、前記着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、前記着目要素の祖先パス名ＩＤをキーとして前記祖先パス出現情報格納部に登録する出現情報登録部とを具備するデータベース構築装置と、
検索式を入力する検索条件入力部と、
前記要素名辞書と前記祖先パス名辞書とを参照して、前記検索条件入力部で入力された検索式について要素名と祖先パス名とをそれぞれ要素名ＩＤと祖先パス名ＩＤとで表現した内部条件式に変換する検索条件解析部と、
前記要素出現情報格納部に記憶している要素出現情報、および、前記祖先パス出現情報格納部に記憶している祖先パス出現情報から、前記検索条件解析部で生成された前記内部条件式にあてはまる検索結果群データを抽出する出現情報取得部とを具備するデータベース検索装置と
を有することを特徴とするデータベース装置。
【請求項１９】
属性名ＩＤと対応する属性名を記憶する属性名辞書と、
前記入力文書解析部の解析結果に基づいて、前記構造化文書に出現する各属性名に対してユニークな属性名ＩＤを割り当てて前記属性名辞書に登録する属性名登録部と、
文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順の情報を少なくとも含む属性出現情報を、属性名ＩＤをキーとして記憶する属性出現情報格納部とをさらに有し、
前記出現情報登録部は、さらに、前記入力文書解析部の解析結果に基づいて、着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順の情報を少なくとも含む属性出現情報を、属性名ＩＤをキーとして前記属性出現情報格納部に登録するようにし、

前記検索条件解析部は、さらに、前記属性名辞書を参照して、前記検索条件入力部で入力された検索式について、属性名を属性ＩＤで表現した内部条件式に変換するようにし、
前記出現情報取得部は、さらに、前記要素出現情報格納部に記憶している要素出現情報と、前記祖先パス出現情報格納部に記憶している祖先パス出現情報と、前記属性出現情報格納部に記憶している属性出現情報とから前記検索条件解析部の出力した前記内部条件式にあてはまる検索結果群データを抽出することを特徴とする請求項１８に記載のデータベース装置。

【書類名】明細書
【発明の名称】データベース構築装置及びデータベース検索装置及びデータベース装置
【技術分野】
【０００１】
本発明は、ＸＭＬなどの論理構造を有する構造化文書を管理するデータベース装置に関し、特に、大量の構造化文書を蓄積管理するデータベース構築装置とそれに蓄積された構造化文書を効率良く検索するデータベース検索装置に関する。
【背景技術】
【０００２】
構造化文書を論理構造に基づいて登録し、論理構造を指定した全文検索をする装置として、構造化文書管理装置が知られている（例えば、特許文献１参照）。
【０００３】
以下、従来例の概要について図を参照しながら説明する。図３３は、従来の構造化文書管理装置の構成図である。登録対象の構造化文書は構造化文書入力部２４０２から入力し、構造解析部２４０７によって解析され、木構造を得る。構造情報作成部２４０８によって、各要素のタグ名（要素名）には名称ＩＤが割り振られて名称ＩＤテーブル格納部２４１８に格納される。また、各要素のパス名称（最上位階層から順にタグ名を連ねて記述した文字列）には、パス名称ＩＤが割り振られて、パス名称インデックス格納部２４１６に格納されている。各要素のパス階層（パス名称の各階層の出現順序（同じ親要素を持つ同じタグ名の要素の中で何番目に出現した要素か）を連ねて記述した文字列）には、パス階層ＩＤが割り当てられて、パス階層インデックス格納部２４１７に格納されている。実体（テキスト）を持つ要素（要素実体）の場合は、各要素実体に対し、検索単位を一意に表す符合（検索単位識別子と呼ぶ）が割り当てられ、この検索単位識別子をキーとして、文書番号、パス名称ＩＤ、パス階層ＩＤ、名称ＩＤの組が要素管理テーブル格納部２４１５に格納される。図３４は、従来の構造化文書管理装置における要素管理テーブルの例を示す図である。図３４においては、要素管理テーブル格納部２４１５に格納される要素管理テーブルの例を示したものである。
【０００４】
次に、文字列索引作成部２４０９は、各要素実体の内容の文字列に対して、予め定めた文字数の文字連鎖を取り出す。この文字連鎖について、該当する検索単位識別子、および該文字連鎖先頭文字がその要素内容において何番目の文字かを表す番号（文字位置番号）を文字列索引格納部２４１９に登録する。図３５は、従来の構造化文書管理装置における文字列索引の例の一部を示す図である。図３５において、２６０１は「検索単位識別子が“１”の要素の文字列中に“構造”という文字連鎖が先頭から“１”文字目の位置から存在する」ということを表している。
【０００５】
次に、このようにして格納されたデータを用いた検索の概要を説明する。図３６は検索条件として「パス名称が“／論文／書誌／タイトル”である要素に“構造化”という文字列が含まれる文書」が与えられた場合の処理を図に示したものである。検索条件解析部２４１０は、パス名称インデックス２４１６を参照し、検索条件のパス名称をパス名称ＩＤ“Ｎ２”に変換する。次に文字列索引検索部２４１１は“構造化”から２文字連鎖“構造”と“造化”を取り出す。文字列索引を参照し、“構造”と“造化”が連続して出現し、かつ検索単位識別子が同一なものを求め、その検索単位識別子を抽出する。図３６は、従来の構造化文書管理装置における検索処理を説明する図である。図３６において、検索単位識別子“１”と“８”が文字列索引検索結果群として返っている。次に、構造照合部２４１２が検索条件の構造指定を満たす最終的な検索結果を求める。文字列索引検索結果群として得られた検索単位識別子をキーにして、要素管理テーブルを参照し、パス名称ＩＤが“Ｎ２”に一致するものだけを最終的な検索結果とする。
【０００６】
その他、タグ名を指定した検索条件であれば、要素管理テーブルの名称ＩＤが指定タグ

名の名称ＩＤと一致するものだけを最終的な検索結果とする。また、パス名称とパス階層をともに指定した検索条件であれば、要素管理テーブルのパス名称ＩＤが指定したパス名称のパス名称ＩＤと一致し、かつパス階層ＩＤが指定したパス階層のパス階層ＩＤと一致するものだけを最終的な検索結果とする。

【０００７】

また、別の文書管理装置として、構造化文書に含まれる要素を階層構造上の位置と結び付けるインデクスを生成し、階層構造上の位置までの探索経路が同じである要素（すなわち１の親ノードに対して複数の子ノードが存在するような構成）であっても複数の要素それぞれを識別するよう管理する文書管理装置が知られている（例えば、特許文献２参照）。

【特許文献１】特開２００２－２０２９７３号公報（第２２頁、第１図）

【特許文献２】特開２００４－３１０６０７号公報（第１４頁、第１図）

【発明の開示】

【発明が解決しようとする課題】

【０００８】

しかしながら、上記従来の構造化文書管理装置では、まず文字列索引を参照して指定された文字列の出現する検索単位識別子を求めた後、検索単位識別子が指定された構造条件を満たすかどうかを、要素管理テーブルを参照して判定するため、文字列検索条件の指定は必須であり、構造条件だけを指定した検索を行うことができない。すなわち、検索を行うためには全ての検索単位識別子について構造条件を満たすかどうかを判定しなければならず、要素管理テーブル全体をサーチしなくてはならないため、効率が非常に悪いという課題がある。また、構造化文書データを蓄積する際に、全文検索のための検索インデクスデータに論理構造データを付加する構造としているため、そのような構造条件だけを指定した検索に対して効率的な検索を可能とする構造の検索用データを構築することができないという課題がある。

【０００９】

また、文字列索引は要素実体の内容文字列に対してのみ作成されるため、要素の属性値に対しては文字列検索を行うことができないという課題がある。

【００１０】

本発明は、このような課題を解決するもので、文字列検索条件と構造条件をともに指定した場合だけでなく、文字列検索条件を伴わない構造だけを指定した様々な検索条件に対しても、所望の文書を効率良く検索することが可能な構造の検索用データを構築し、効率良く検索可能なデータベース装置を提供することを目的とする。

【００１１】

また、本発明は、要素内のテキスト文字列だけでなく、属性値に対しても文字列検索が可能な検索用データを構築し、効率良く検索可能なデータベース装置を提供することを目的とする。

【課題を解決するための手段】

【００１２】

前記従来の課題を解決するために、本発明のデータベース構築装置は、構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行う入力文書解析部と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名ＩＤを割り当てて要素名辞書に登録する要素名登録部と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて祖先パス名辞書に登録する祖先パス名登録部と、入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして要素出現情報格納部に登録し、かつ、文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして祖先パス出現情報格納部に登録する出現情報登録部とを備える。

【００１３】

そのため、構造化文書を登録蓄積する際に、要素の出現情報に基づいて適切な出現情報インデクスを生成し、文字列検索条件と構造条件をともに指定した場合だけでなく、文字列検索条件を伴わない構造条件だけを指定した様々な検索条件に対しても、所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

【００１４】

また、本発明のデータベース構築装置は、入力文書解析部の解析結果に基づいて、構造化文書に出現する各属性名に対してユニークな属性名ＩＤを割り当てて属性名辞書に登録する属性名登録部を有し、出現情報登録部が、入力文書解析部の解析結果に基づいて、着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順の情報を少なくとも含む属性出現情報を、属性名ＩＤをキーとして属性出現情報格納部に登録する。

【００１５】

そのため、構造化文書の登録の際に、属性に関する構造情報を登録できるようになり、結果として属性に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

【００１６】

また、本発明のデータベース構築装置は、出現情報登録部が、入力文書解析部の解析結果に基づいて、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと属性名ＩＤと分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納部に登録する。

【００１７】

そのため、構造化文書の登録の際に、要素実体テキストおよび属性値の部分文字列に関する構造情報を登録できるようになり、結果として要素実体テキストおよび属性値の部分文字列に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

【００１８】

また、本発明のデータベース構築装置は、要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順と空要素順の情報を少なくとも含み、祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順と空要素順の情報を少なくとも含む。

【００１９】

そのため、構造化文書の登録の際に、要素が要素実体のテキストを全く含まない要素（空要素）に関する構造情報を登録できるようになり、結果として空要素に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

【００２０】

また、本発明のデータベース構築装置は、要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順と空要素順の情報を少なくとも含み、祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順と空要素順の情報を少なくとも含み、属性出現情報は、着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順と空要素順の情報を少なくとも含む。

【００２１】

そのため、構造化文書の登録の際に、属性がテキストを全く含まない要素（空要素）に関する構造情報を登録できるようになり、結果として属性の空要素に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

【００２２】

また、本発明のデータベース構築装置は、要素出現情報は、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順と空要素順の情報を少なくとも含み、祖先パス出現情報は、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順と空要素順の情

報を少なくとも含み、属性出現情報は、着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順と空要素順の情報を少なくとも含み、テキスト出現情報は、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと属性名ＩＤと分岐順と空要素順の情報を少なくとも含む。

【００２３】

そのため、構造化文書の登録の際に、要素実体テキストおよび属性値から切り出された部分文字列がテキストを全く含まない要素（空要素）に関する構造情報を登録できるようになり、結果として要素実体テキストおよび属性値から切り出された部分文字列の空要素に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

【００２４】

また、本発明のデータベース構築装置は、祖先パス名登録部は、構造化文書に出現する各祖先パス名を１つ以上に分割した各々の部分祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて祖先パス名辞書に登録する。

【００２５】

そのため、構造化文書の登録の際に、祖先パス名を分割して部分パスを重複して蓄積しないように祖先パス列として登録できるようになり、結果として祖先パス辞書のサイズが小さく、構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

【００２６】

また、本発明のデータベース構築装置は、要素出現情報格納部に同じ要素名ＩＤをキーにして登録されている要素出現情報のエントリ群と、祖先パス出現情報格納部に同じ祖先パス名ＩＤをキーにして登録されている祖先パス出現情報のエントリ群とに対して、文書番号と文字位置以外の１つ以上の情報項目の値が共通するエントリ同士をグループ化する出現情報グループ化部を備える。

【００２７】

そのため、登録されている構造化文書の出現位置情報の共通する値の項目を重複して蓄積しないようにグループ化して登録できるようになり、結果として出現位置索引のサイズが小さく、構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができる。

【００２８】

また、本発明のデータベース検索装置は、構造化文書に出現する各要素名に対してユニークな要素名ＩＤを登録した要素名辞書と、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを登録した祖先パス名辞書と、構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして格納した要素出現情報格納部と、構造化文書の解析結果に基づいて、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして格納した、祖先パス出現情報格納部と、検索式を入力するための検索条件入力部と、要素名辞書と祖先パス名辞書とを参照して、入力された検索式を内部条件式に変換する検索条件解析部と、検索条件解析部の出力した内部条件式にしたがって、要素出現情報格納部からの要素出現情報および、祖先パス出現情報格納部からの祖先パス出現情報から検索結果群を求める出現情報取得部とを備える。

【００２９】

そのため、構造化文書を検索する際に、要素と祖先パスの出現情報に基づく適切な出現情報インデクスを参照できるようになり、結果として文字列検索条件を伴わない要素名と祖先パス名に関する構造条件だけを指定した検索条件に対して所望の構造化文書を効率良く検索することができる。

【００３０】

また、本発明のデータベース検索装置は、属性名ＩＤと対応する属性名の記録された属性名辞書と、着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順の情報を少なくとも含む属性出現情報を、属性名ＩＤをキーとして格納した属性出現情報格納部とを有し、検索条件解析部が、要素名辞書と祖先パス名辞書と属性名辞書とを参照して、検索条件入力部から入力された検索式を内部条件式に変換し、出現情報取得部が、検索条件解析部の出力した内部条件式にしたがって、要素出現情報格納部からの要素出現情報、祖先パス出現情報格納部からの祖先パス出現情報および、属性出現情報格納部からの属性出現情報から検索結果群を求める。

【００３１】

そのため、構造化文書を検索する際に、要素名と祖先パス名と属性名に関する出現情報インデクスを参照できるようになり、結果としてそれらに関する構造条件だけを指定した検索条件に対して所望の構造化文書を効率良く検索することができる。

【００３２】

また、本発明のデータベース検索装置は、要素実体テキストおよび属性値から切り出された部分文字列に関し、出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと属性名ＩＤと分岐順の情報を少なくとも含むテキスト出現情報を、切り出された部分文字列をキーとして格納した、テキスト出現情報格納部とを有し、出現情報取得部が、検索条件解析部の出力した内部条件式にしたがって、要素出現情報格納部からの要素出現情報、祖先パス出現情報格納部からの祖先パス出現情報、属性出現情報格納部からの属性出現情報および、テキスト出現情報格納部からのテキスト出現情報から検索結果群を求める。

【００３３】

そのため、構造化文書を検索する際に、要素名と祖先パス名と属性名と要素実体テキストおよび属性値から切り出された部分文字列に関する出現情報インデクスを参照できるようになり、結果としてそれらに関する構造条件だけを指定した検索条件に対して所望の構造化文書を効率良く検索することができる。

【００３４】

また、本発明のデータベース検索装置は、出現情報取得部は、要素出現情報格納部における指定要素名ＩＤのエントリ数と、祖先パス出現情報格納部における指定祖先パス名ＩＤのエントリ数の大小を比較し、少ない方の出現情報を参照するようにして検索結果群を求める。

【００３５】

そのため、構造化文書を検索する際に、構造化文書に含まれる論理構造の要素数に応じて少ないエントリの出現情報を選択できるようになり、結果として検索対象が出現するエントリ数の絞込みが速く、構造条件だけを指定した検索条件に対して所望の構造化文書を効率良く検索することができる。

【００３６】

また、本発明のデータベース装置は、構造化文書に出現する各要素名に対してユニークな要素名ＩＤを記憶する要素名辞書と、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを記憶する祖先パス名辞書と、構造化文書にユニークな文書番号を割り当てるとともに構造の解析を行う入力文書解析部と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各要素名に対してユニークな要素名ＩＤを割り当てて要素名辞書に登録する要素名登録部と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて祖先パス名辞書に登録する祖先パス名登録部と、文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、要素名ＩＤをキーとして記憶する要素出現情報格納部と、文書番号と文字位置と要素名ＩＤと分岐順の情報を少なくとも含む祖先パス出現情報を、祖先パス名ＩＤをキーとして記憶する祖先パス出現情報格納部と、入力文書解析部の解析結果に基づいて、着目要素の出現する文書番号と文字位置と祖先パス名ＩＤと分岐順の情報を少なくとも含む要素出現情報を、着目要素の要素名ＩＤをキーとして要素出現情報格納部に登録し、かつ、着目要素の出現する文書番号と文字位置と要素名ＩＤと分岐順の情

報を少なくとも含む祖先パス出現情報を、着目要素の祖先パス名ＩＤをキーとして祖先パス出現情報格納部に登録する出現情報登録部とを具備するデータベース構築装置と、検索式を入力する検索条件入力部と、要素名辞書と祖先パス名辞書とを参照して、検索条件入力部で入力された検索式について要素名と祖先パス名とをそれぞれ要素名ＩＤと祖先パス名ＩＤとで表現した内部条件式に変換する検索条件解析部と、要素出現情報格納部に記憶している要素出現情報、および、祖先パス出現情報格納部に記憶している祖先パス出現情報から、検索条件解析部で生成された内部条件式にあてはまる検索結果群データを抽出する出現情報取得部とを具備するデータベース検索装置とを備える。

【００３７】

そのため、要素の出現情報に基づいて適切な出現情報インデクスを生成し、文字列検索条件と構造条件をともに指定した場合だけでなく、文字列検索条件を伴わない構造条件だけを指定した様々な検索条件に対しても、所望の文書を効率良く検索することが可能な構造の検索用データを構築し、また、効率良く検索することができる。

【００３８】

また、本発明のデータベース装置は、属性名ＩＤと対応する属性名を記憶する属性名辞書と、入力文書解析部の解析結果に基づいて、構造化文書に出現する各属性名に対してユニークな属性名ＩＤを割り当てて属性名辞書に登録する属性名登録部と、文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順の情報を少なくとも含む属性出現情報を、属性名ＩＤをキーとして記憶する属性出現情報格納部とをさらに有し、出現情報登録部は、さらに、入力文書解析部の解析結果に基づいて、着目属性の出現する文書番号と文字位置と祖先パス名ＩＤと要素名ＩＤと分岐順の情報を少なくとも含む属性出現情報を、属性名ＩＤをキーとして属性出現情報格納部に登録するようにし、検索条件解析部は、さらに、属性名辞書を参照して、検索条件入力部で入力された検索式について、属性名を属性ＩＤで表現した内部条件式に変換するようにし、出現情報取得部は、さらに、要素出現情報格納部に記憶している要素出現情報と、祖先パス出現情報格納部に記憶している祖先パス出現情報と、属性出現情報格納部に記憶している属性出現情報とから検索条件解析部の出力した内部条件式にあてはまる検索結果群データを抽出する。

【００３９】

そのため、構造化文書の登録の際に、属性に関する構造情報を登録できるようになり、結果として属性に関する構造条件を指定して所望の文書を効率良く検索することが可能な構造の検索用データを構築することができ、また、効率良く検索することができる。

【発明の効果】

【００４０】

本発明のデータベース装置によれば、文字列検索条件と構造条件をともに指定した検索条件のみならず、構造だけを指定した様々な検索条件に対しても、所望の論理構造を持つ文書を効率良く検索するデータベースが構築でき、さらに効率良く検索することが可能となる。

【００４１】

また、要素実体のテキスト文字列に対してだけでなく、属性値に対しても文字列検索を行うことが可能となる。

【発明を実施するための最良の形態】

【００４２】

以下、本発明の実施の形態におけるデータベース装置について、図面を参照しながら説明する。

【００４３】

（実施の形態１）

本実施の形態におけるデータベース装置の構成および動作について説明する。図１は、本発明の実施の形態１におけるデータベース装置の構成を示すブロック図である。図１において、１０１はデータベースに登録する構造化文書群、１０２は入力された構造化文書群１０１の各文書についてユニークな文書番号を割り振るとともに論理構造の解析を行う

入力文書解析部、１０３は入力文書解析部１０２の解析結果から、文書に出現する要素名に対してユニークな識別子（以下、要素名ＩＤと呼ぶ）を割り当てて要素名辞書１０７に登録する要素名登録部、１０４は入力文書解析部１０２の解析結果から、文書に出現する祖先パス名（着目要素の祖先要素の要素名を最上位階層から順にスラッシュで区切って並べた文字列で、着目要素自身の要素名は含まない）に対してユニークな識別子（以下、祖先パス名ＩＤと呼ぶ）を割り当てて祖先パス名辞書１０８に登録する祖先パス名登録部、１０５は入力文書解析部１０２の解析結果から、文書に出現する属性名に対してユニークな識別子（以下、属性名ＩＤと呼ぶ）を割り当てて属性名辞書１０９に登録する属性名登録部、１０６は入力文書解析部１０２の解析結果から、出現位置索引１１０の要素出現情報格納部１１１、祖先パス出現情報格納部１１２、属性出現情報格納部１１３、テキスト出現情報格納部１１４に４種の出現情報を登録する出現情報登録部、１０７は要素名ＩＤとそれに対応する要素名が記録された要素名辞書、１０８は祖先パス名ＩＤとそれに対応する祖先パス名が記録された祖先パス名辞書、１０９は属性名ＩＤとそれに対応する属性名が記録された属性名辞書、１１０は要素出現情報格納部１１１、祖先パス出現情報格納部１１２、属性出現情報格納部１１３、テキスト出現情報格納部１１４、の４種の出現情報が格納されている出現位置索引格納部、１１１は各要素の出現する文書番号、文字位置、文字数、祖先パス名ＩＤ、分岐順の情報を、要素名ＩＤをキーにして格納した要素出現情報格納部、１１２は各要素の出現する文書番号、文字位置、文字数、要素名ＩＤ、分岐順の情報を、その要素の祖先パス名ＩＤをキーにして格納した、祖先パス出現情報格納部、１１３は各属性の出現する文書番号、文字位置、文字数、要素名ＩＤ、祖先パス名ＩＤ、分岐順の情報を、属性名ＩＤをキーにして格納した属性出現情報格納部、１１４は要素内のテキストから切り出した部分文字列、および要素の持つ属性の値から切り出した部分文字列に関して、出現する文書番号、文字位置、祖先パス名ＩＤ、要素名ＩＤ、属性名ＩＤ、分岐順の情報を、部分文字列をキーにして格納したテキスト出現情報格納部、１１６は検索式１１５を受け付ける検索条件入力部、１１７は、検索条件入力部１１６に与えられた検索式を解析し、内部条件に変換して出現情報取得部１１８に出力する検索条件解析部、１１８は検索条件解析部１１７の出力した内部条件にしたがって、出現位置索引１１０に格納された４種の出現情報から適切な情報を選択して取得し、検索条件にマッチする結果データ集合を求める出現情報取得部、１１９は結果データ集合を適切な形式で検索結果１２０として出力する検索結果出力部である。

【００４４】

上記のように構成されたデータベース装置の動作について説明する。はじめに、文書登録（データベース構築）処理に関して具体例を挙げて説明する。図２は、本発明の実施の形態１における文書登録処理の手順を示す流れ図である。

【００４５】

まず、ステップ２２０１において、入力文書解析部１０２は、構造化文書群１０１から構造化文書を１つ読み込んで、ユニークな文書番号を割り振る。

【００４６】

次に、ステップ２２０２において、入力文書解析部１０２は、この文書の論理構造を解析する。図３は、本発明の実施の形態１における登録検索対象となる構造化文書の一例を示す図である。構造化文書群１０１には、このような図３に示す文書が複数含まれる。図３に示した構造化文書は、最上位階層にｂｏｏｋ要素を持ち、ｂｏｏｋ要素はｔｉｔｌｅ要素と２つのｃｈａｐｔｅｒ要素を含んでいる。ｔｉｔｌｅ要素は、要素実体の文字列“文書検索”を含み、１つ目のｃｈａｐｔｅｒ要素は別のｔｉｔｌｅ要素と２つのｓｅｃｔｉｏｎ要素および属性値が“歴史”であるｋｅｙｗｏｒｄ属性を持つ構造を持っている。図３に示す構造化文書を入力文書解析部１０２によって解析した結果得られる木構造は、図４のようになる。図４は、本発明の実施の形態１における構造化文書の論理構造を解析した結果である木構造の一例を示す図である。図４において、四角い枠は要素３０１～３０３を表し、枠内に記された文字列は要素名３０４を示している。また、楕円の点線枠は属性３０５を表し、枠内に記された文字列は属性名３０６を示している。木構造の最上位

階層の要素３０１から着目要素に至る経路の途中に存在する要素（祖先要素）の要素名をスラッシュで区切って順に並べたものはパス名と呼ばれる。パス名のうちの末尾部分（＝着目要素自身の要素名）を除いた部分を「祖先パス名」と呼ぶことにする。図５は、本発明の実施の形態１における祖先パス名を説明する図である。図５において、図４の網掛けを施した要素３０２に関するパス名７０１、祖先パス名７０２、要素名７０３を示している。

【００４７】

また、図４において、要素の右肩に記された“１／２／３”などの文字列は、パス名中の各要素について、同じ親要素を持つ同じ要素名の要素の中で何番目に出現したかの順を示す番号を並べたもので、これを「分岐順」３０７と呼ぶ。図４の網掛けを施した要素３０２とその左隣の要素３０３とは、パス名は同じであるが分岐順３０７、３０８は異なっている。なお、分岐順の表記方法はこれに限らない。例えば、１以外の値を持つ階層の深さとその値を並べる方法でもよい。分岐順３０７（“１／２／３”）をこの方法で表記すれば、深さ１の値は１なので省略、深さ２の値が２、深さ３の値が３、したがって“２：２，３：３”となる。同じ要素名の兄弟要素がめったに現れない文書、すなわち、分岐順の値がほとんど１であるような文書を格納する場合には、このような表記方法の方が出現位置索引ファイルのサイズを小さくできる。

【００４８】

次に、入力文書解析部１０２の解析結果をうけて、当該文書に出現する各要素について以下の処理を繰り返す。

【００４９】

ステップ２２０３において、要素名登録部１０３は、着目要素の要素名が要素名辞書１０７に登録済みかどうかを調べ、登録済みであれば対応する要素名ＩＤを取得し、登録されていなければ新たに要素名ＩＤ（＞０）を割り当てて要素名辞書１０７に登録する。

【００５０】

ステップ２２０４において、祖先パス名登録部１０４は、着目要素の祖先パス名が祖先パス名辞書１０８に登録済みかどうかを調べ、登録済みであれば対応する祖先パス名ＩＤを取得し、登録されていなければ新たに祖先パス名ＩＤ（＞０）を割り当てて祖先パス名辞書１０８に登録する。

【００５１】

もし、着目要素が属性を持っているならば、ステップ２２０５～ステップ２２０６において、属性名登録部１０５は、着目要素の各属性の属性名が属性名辞書１０９に登録済みかどうかを調べ、登録済みであれば対応する属性名ＩＤを取得し、登録されていなければ新たに属性名ＩＤ（＞０）を割り当てて属性名辞書１０９に登録する。図６は、本発明の実施の形態１における要素名辞書の内容の一例を示す図である。また、図７は、本発明の実施の形態１における祖先パス名辞書の内容の一例を示す図である。また、図８は、本発明の実施の形態１における属性名辞書の内容の一例を示す図である。図７、図８、図９において、それぞれ構造化文書（図３）の登録処理が終わった後の要素名辞書１０７、祖先パス名辞書１０８、属性名辞書１０９の内容の例を示している。

【００５２】

ステップ２２０７において、出現情報登録部１０６は、着目要素に関する要素出現情報を、要素名ＩＤをキーとして要素出現情報格納部１１１に登録する。要素出現情報は、文書番号、着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置および文字数、祖先パス名ＩＤ、分岐順の５種類の値の組から構成される。なお、「文字位置」は、図９に示すように、タグを除く当該文書内の全てのテキストをつなげた文字列において先頭から何文字目にあたるかで表す。また、着目要素が要素実体のテキストを全く含まない要素（＝空要素）である場合には、着目要素以降に初めて現れる（タグ以外の）テキストの先頭文字位置を着目要素の先頭文字位置とみなす。図１０は、本発明の実施の形態１における要素出現情報を説明する図である。図１０において、図４の網掛けを施した要素３０２に関する要素出現情報が、要素名ＩＤが４（＝要素名がｓｅｃｔｉｏｎ）

である要素が文書番号１の文書の１１５文字目から始まる長さ４０文字の要素実体を含んでいて、その祖先パス名ＩＤが３（＝祖先パス名が／ｂｏｏｋ／ｃｈａｐｔｅｒ）で分岐順が１／２／３であることを表している。

【００５３】

ステップ２２０８において、出現情報登録部１０６は、着目要素に関する祖先パス出現情報（すなわち、文書番号、着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置および文字数、要素名ＩＤ、分岐順の５種類の値の組）を、祖先パス名ＩＤをキーとして祖先パス出現情報格納部１１２に登録する。図１１は、本発明の実施の形態１における祖先パス出現情報を説明する図である。図１１において、図４の網掛けを施した要素３０２に関する祖先パス出現情報の内容を示している。図１０と図１１を比較してわかるように、同一要素に関する要素出現情報と祖先パス出現情報は、キーとなる項目が要素名ＩＤであるか祖先パス名ＩＤであるかという点が異なるだけである。

【００５４】

もし、着目要素が属性を持っているならば、ステップ２２０９～ステップ２２１０において、出現情報登録部１０６は着目要素の各属性に関する属性出現情報を、属性名ＩＤをキーとして属性出現情報格納部１１３に登録する。属性出現情報は、文書番号、属性値の先頭文字位置および文字数、祖先パス名ＩＤ、要素名ＩＤ、分岐順の６種類の値の組から構成される。図１２は、本発明の実施の形態１における属性出現情報を説明する図である。図１２において、図４の網掛けを施した要素３０２の「ｕｐｄａｔｅ」属性３０５に関する属性出現情報の内容を示している。その内容は、属性名ＩＤが２（＝属性名がｕｐｄａｔｅ）の属性が文書番号１の文書の１１５文字目から始まる長さ６文字の属性値を持ち、属性の所属する要素の祖先パス名ＩＤが３（＝祖先パス名が／ｂｏｏｋ／ｓｅｃｔｉｏｎ）、要素名ＩＤが４（＝要素名がｓｅｃｔｉｏｎ）、分岐順が１／２／３であることを示している。なお、属性出現情報において、属性値の先頭文字位置は、図１２に示すように、仮想的に着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置と同じであるとする。

【００５５】

ステップ２２１１において、出現情報登録部１０６は、着目要素の実体内容のテキストから部分文字列の切り出しを行い、テキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納部１１４に登録する。ただし、属性値ではないので、属性名ＩＤには常に０を格納する。テキスト出現情報は、文書番号、切り出された部分文字列の先頭文字位置、祖先パス名ＩＤ、要素名ＩＤ、属性名ＩＤ、分岐順の６種類の値の組から構成される。

【００５６】

もし、着目要素が属性を持っているならば、ステップ２２１２～ステップ２２１３において、出現情報登録部１０６は、着目要素が持つ各属性の属性値文字列から部分文字列の切り出しを行い、テキスト出現情報格納部１１４に部分文字列をキーとして登録する。なお、属性出現情報と同様に、属性値は図１１に示すような位置に仮想的に出現しているとして、文字位置を算出する。また、ステップ２２１３ではステップ２２１１の場合とは異なり、属性名ＩＤには着目している属性の属性名ＩＤ（＞０）を格納する。図１３は、本発明の実施の形態１におけるテキスト出現情報を説明する図である。図１３において、図４の網掛けを施した要素３０２のテキストおよび「ｕｐｄａｔｅ」属性３０５の属性値についてのテキスト出現情報の一部である。図１３において、１２０１は、“極大”という部分文字列が文書番号１の文書の１１８文字目に現れ、祖先パス名ＩＤが３（＝祖先パス名が／ｂｏｏｋ／ｓｅｃｔｉｏｎ）、要素名ＩＤが４（要素名がｃｈａｐｔｅｒ）、分岐順が１／２／３であるような要素の要素実体に含まれている（属性名ＩＤが０であることからわかる）ことを表している。また１２０２は、“００”という部分文字列が文書番号１の文書の１１６文字目に現れ、祖先パス名ＩＤが３（＝祖先パス名が／ｂｏｏｋ／ｓｅｃｔｉｏｎ）、要素名ＩＤが４（＝要素名がｃｈａｐｔｅｒ）、分岐順が１／２／３であるような要素に属する属性名ＩＤが２（＝属性名がｕｐｄａｔｅ）の属性の属性値に含ま

れていることを表している。

【００５７】

ステップ２２１４において、この文書に出現する全ての要素について処理が終わったかどうかを調べ、もし未処理の要素が残っていればステップ２２０３に戻って処理を繰り返す。

【００５８】

ステップ２２１５において、全ての入力文書に対して処理が終わったかどうかを調べ、未処理の文書が残っていればステップ２２０１に戻って処理を繰り返す。

【００５９】

以上のようにして、文書登録（データベース構築）処理が完了する。

【００６０】

続いて、登録済みの文書群に対する検索処理に関して説明する。図１４は、本発明の実施の形態１における検索式の例を示す図である。図１４においては、検索条件入力部１１６に与えられる検索式１１５の例をいくつか示したもので、これらの式はＷ３Ｃ（Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂ　Ｃｏｎｓｏｒｔｉｕｍ）の勧告として公開されているＸＰａｔｈ言語（詳細な仕様はｈｔｔｐ：／／ｗｗｗ．ｗ３．ｏｒｇ／ＴＲ／ｘｐａｔｈに記載されている）で記述されている。

【００６１】

図１４のそれぞれのＸＰａｔｈ式は、次のような意味を表している。検索式２１０１は「最上位階層のｂｏｏｋ要素の子のｃｈａｐｔｅｒ要素の子であるｔｉｔｌｅ要素」を表している。検索式２１０２は「最上位階層のｂｏｏｋ要素の子のｃｈａｐｔｅｒ要素のいずれかの子要素」を表している。検索式２１０３は、「いずれかの階層にあるｔｉｔｌｅ要素」を表している。検索式２１０４は「最上位階層のｂｏｏｋ要素の子のｃｈａｐｔｅｒ要素の子の２番目のｓｅｃｔｉｏｎ要素」を表している。検索式２１０５は、「最上位階層のｂｏｏｋ要素の子のｃｈａｐｔｅｒ要素の子のｓｅｃｔｉｏｎ要素のｕｐｄａｔｅ属性」を表している。検索式２１０６は、「最上位階層のｂｏｏｋ要素の子のｃｈａｐｔｅｒ要素の子のｓｅｃｔｉｏｎ要素で、かつ要素実体内容に“極大単語”という文字列を含む要素」を表している。検索式２１０７は、「最上位階層のｂｏｏｋ要素の子のｃｈａｐｔｅｒ要素の子のｓｅｃｔｉｏｎ要素のｕｐｄａｔｅ属性で、かつその属性値に“２００４”という文字列を含む」を表している。

【００６２】

次に、それぞれの検索式に対して、本実施の形態におけるデータベース装置でどのような検索処理が行われるのかを順に説明する。図１５は、本発明の実施の形態１におけるデータベース装置の検索処理の手順を示す流れ図である。

【００６３】

（検索式２１０１の場合）

図１５に沿って、検索式２１０１の場合の検索処理の流れを説明する。

ステップ２３０１において、検索条件入力部１１６に入力された検索式２１０１は、検索条件解析部１１７で解析される。

【００６４】

ステップ２３０２において、検索条件解析部１１７は、検索式２１０１を解析し、要素名辞書１０７、祖先パス名辞書１０８を参照して内部条件「祖先パス名ＩＤ＝３かつ要素名ＩＤ＝２」に変換し、出現情報取得部１１８に出力する。

【００６５】

次に、ステップ２３０３からステップ２３０５において、出現情報取得部１１８は、出現位置索引１１０を参照し、要素出現情報格納部１１１における要素名ＩＤ＝２のエントリ数Ｎと祖先パス出現情報格納部１１２における祖先パス名ＩＤ＝３のエントリ数Ｍとを比較し、少ない方を選択する。図１６は、要素出現情報格納部１１１における要素名ＩＤ＝２のエントリ１３０１、図１７は祖先パス出現情報格納部１１２における祖先パス名ＩＤ＝３のエントリ１４０１の例で、この場合はＮ＝８、Ｍ＝１２であるから図１６の要素

出現情報格納部１１１を選ぶことになる。
【００６６】
そして、ステップ２３０６において、出現情報取得部１１８は、要素出現情報格納部１１１の要素名ＩＤ＝２のエントリ１３０１から１つ取得し、ステップ２３０７で、このエントリの祖先パス名ＩＤが３であるかどうかを調べ、もし祖先パス名ＩＤが３であればステップ２３０８でこのエントリのデータを結果データ集合１３０２に追加する。結果データ集合の各データは例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式である。
【００６７】
ステップ２３０９において、出現情報取得部１１８は、Ｎエントリ全てについて処理したか調べ、まだ未処理のエントリがあればステップ２３０６に戻って処理を繰り返す。
【００６８】
ステップ２３０５において、出現情報取得部１１８は、もしＭ≦Ｎであれば、図１７のように祖先パス出現情報格納部１１２における祖先パス名ＩＤ＝３の各エントリ１４０１を調べ、要素名ＩＤが２であるものを求め（ステップ２３１０～ステップ２３１３）結果データ集合１４０２に追加する。
【００６９】
ステップ２３１４において、出現情報取得部１１８は、求められた結果データ集合を検索結果出力部１１９に出力する。
【００７０】
最後に検索結果出力部１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。
【００７１】
このように、検索式２１０１に対しては、要素出現情報格納部１１１における指定した要素名ＩＤのエントリから指定した祖先パス名ＩＤを持つものを選ぶという処理と、祖先パス出現情報格納部１１２における指定した祖先パス名ＩＤのエントリから指定した要素名ＩＤを持つものを選ぶという２種類の処理のどちらか、エントリ数の少ない方を選ぶことによって、検索対象構造化文書群の論理構造の特性に応じて処理量を抑えることができるため、所望の文書を効率良く検索することができる。
【００７２】
（検索式２１０２の場合）
検索条件入力部１１６に入力された検索式２１０２は、検索条件解析部１１７で解析される。検索条件解析部１１７は、検索式２１０２を解析し、祖先パス名辞書１０８を参照して内部条件「祖先パス名ＩＤ＝３」に変換し、出現情報取得部１１８に出力する。出現情報取得部１１８は、出現位置索引１１０を参照し、図１８のように祖先パス出現情報格納部１１２における祖先パス名ＩＤ＝３の全てのエントリ１５０１を求め、例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で結果データ集合１５０２として検索結果出力部１１９に出力する。検索結果出力部１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。
【００７３】
このように、検索式２１０２に対しては、祖先パス出現情報格納部１１２における指定した祖先パス名ＩＤのエントリを取得するだけで良いため、所望の文書を効率良く検索することができる。
【００７４】
（検索式２１０３の場合）
検索条件入力部１１６に入力された検索式２１０３は、検索条件解析部１１７で解析される。検索条件解析部１１７は、検索式２１０３を解析し、要素名辞書１０７を参照して内部条件「要素名ＩＤ＝２」に変換し、出現情報取得部１１８に出力する。出現情報取得部１１８は、出現位置索引１１０を参照し、図１９のように要素出現情報格納部１１１における要素名ＩＤ＝２の全てのエントリ１６０１を求め、例えば（文書番号，祖先パス名

ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で結果データ集合１６０２を検索結果出力部１１９に出力する。検索結果出力部１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【００７５】

このように、検索式２１０３に対しては、要素出現情報格納部１１１における指定した要素名ＩＤのエントリを取得するだけで良いため、所望の文書を効率良く検索することができる。

【００７６】

（検索式２１０４の場合）

検索条件入力部１１６に入力された検索式２１０４は、検索条件解析部１１７で解析される。検索条件解析部１１７は、検索式２１０４を解析し、要素名辞書１０７、祖先パス名辞書１０８を参照して内部条件「祖先パス名ＩＤ＝３かつ要素名ＩＤ＝４かつ分岐順＝”＊／＊／２”」に変換し、出現情報取得部１１８に出力する。分岐順のアスタリスク「＊」の部分はどんな数字でもマッチすることを表す。出現情報取得部１１８は、出現位置索引１１０を参照し、要素出現情報格納部１１１における要素名ＩＤ＝４のエントリ数Ｎと祖先パス出現情報格納部１１２における祖先パス名ＩＤ＝３のエントリ数Ｍとを比較し、少ない方を選択する。

【００７７】

もし、Ｍ≦Ｎであれば、図２０に示すように祖先パス出現情報格納部１１２における祖先パス名ＩＤ＝３の各エントリ１７０１を調べ、要素名ＩＤが４であり、かつ分岐順が”＊／＊／２”であるエントリのデータを結果データ集合１７０２として、例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で検索結果出力部１１９に出力する。もし、Ｍ＞Ｎならば要素出現情報格納部１１１における要素名ＩＤ＝４の各エントリを調べ、祖先パス名ＩＤが３であり、かつ分岐順が“＊／＊／２”であるエントリのデータを結果データ集合１７０２として検索結果出力部１１９に出力する。

【００７８】

最後に検索結果出力部１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【００７９】

このように、検索式２１０４に対しては、要素出現情報格納部１１１における指定した要素名ＩＤのエントリから指定した祖先パス名ＩＤと分岐順を持つものを選ぶという処理と、祖先パス出現情報格納部１１２における指定した祖先パス名ＩＤのエントリから指定した要素名ＩＤと分岐順を持つものを選ぶという２種類の処理のどちらか、エントリ数の少ない方を選ぶ。このことによって、処理量を減らすことが可能となり、所望の文書を効率良く検索することができる。

【００８０】

（検索式２１０５の場合）

検索条件入力部１１６に入力された検索式２１０５は、検索条件解析部１１７で解析される。検索条件解析部１１７は、検索式２１０５を解析し、要素名辞書１０７、祖先パス名辞書１０８、属性名辞書１０９を参照して内部条件「祖先パス名ＩＤ＝３かつ要素名ＩＤ＝４かつ属性名ＩＤ＝２」に変換し、出現情報取得部１１８に出力する。出現情報取得部１１８は、出現位置索引１１０を参照し、図２１のように属性出現情報格納部１１３における属性名ＩＤ＝２の各エントリ１８０１を調べ、祖先パス名ＩＤが３であり、要素名ＩＤが４であればそのエントリのデータを例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で結果データ集合１８０２として検索結果出力部１１９に出力する。最後に、検索結果出力部１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【００８１】

このように、検索式２１０５に対しては、属性出現情報格納部１１３における指定した属性名ＩＤのエントリから指定した祖先パス名ＩＤと要素名ＩＤを持つものを選ぶことに

よって、所望の文書を検索することが可能となる。

【００８２】

（検索式２１０６の場合）

検索条件入力部１１６に入力された検索式２１０６は、検索条件解析部１１７で解析される。検索条件解析部１１７は、検索式２１０６を解析し、要素名辞書１０７、祖先パス名辞書１０８を参照して内部条件「祖先パス名ＩＤ＝３かつ要素名ＩＤ＝４かつ要素内に文字列“極大単語”を含む」に変換し、出現情報取得部１１８に出力する。出現情報取得部１１８は、出現位置索引１１０を参照し、図２２のようにテキスト出現情報格納部１１４における“極大”のエントリ１９０１と“単語”のエントリ１９０２の間の連接演算を行う。その際、文書番号が同一であることと“単語”が“極大”の２文字後方に位置することだけでなく、祖先パス名ＩＤが３、かつ要素名ＩＤが４、かつ属性名ＩＤが０、かつ分岐順が同一であるというチェックも行い条件を満たすものを出力する。例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で結果データ集合１９０３として検索結果出力部１１９に出力する。検索結果出力部１１９は、求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【００８３】

このように、検索式２１０６に対しては、テキスト出現情報格納部１１４における部分文字列のエントリ同士の連接演算の際に、祖先パス名ＩＤおよび要素名ＩＤが指定した値であって、分岐順が同一であり、かつ属性名ＩＤが０であるものを選ぶことによって、所望の文書を検索することが可能となる。

【００８４】

（検索式２１０７の場合）

検索条件入力部１１６に入力された検索式２１０７は、検索条件解析部１１７で解析される。検索条件解析部１１７は、検索式２１０７を解析し、要素名辞書１０７、祖先パス名辞書１０８、属性名辞書１０９を参照して内部条件「祖先パス名ＩＤ＝３かつ要素名ＩＤ＝４かつ属性名ＩＤ＝２かつ属性値に文字列“２００４”を含む」に変換し、出現情報取得部１１８に出力する。出現情報取得部１１８は、出現位置索引１１０を参照し、図２３のようにテキスト出現情報格納部１１４における“２０”のエントリ２００１と“０４”のエントリ２００２の間の連接演算を行う。その際、文書番号が同一であることと“２０”が“０４”の２文字後方に位置することだけでなく、祖先パス名ＩＤが３、かつ要素名ＩＤが４、かつ属性名ＩＤが２、かつ分岐順が同一であるというチェックも行い、条件を満たすものを出力する。例えば（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順）のような形式で結果データ集合２００３として検索結果出力部１１９に出力する。検索結果出力部１１９は求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【００８５】

このように、検索式２１０７に対しては、テキスト出現情報格納部１１４における部分文字列のエントリ同士の連接演算の際に、祖先パス名ＩＤおよび要素名ＩＤが指定した値であって、分岐順が同一であり、かつ属性名ＩＤが指定した値（＞０）であるものを選ぶことによって、所望の文書を検索することが可能となる。

【００８６】

以上説明したように、要素の出現情報を、要素名ＩＤをキーにして格納した要素出現情報格納部と、要素の出現情報をその要素の祖先パス名ＩＤをキーにして格納した祖先パス出現情報格納部と、属性の出現情報を、属性名ＩＤをキーにして格納した属性出現情報格納部とを設けることにより、構造条件だけを指定した検索式に対しても効率良く所望の文書を検索することができる。また、要素実体のテキスト文字列および要素の持つ属性の属性値から切り出された部分文字列の出現情報を格納したテキスト出現情報格納部を設けることにより、要素実体のテキストに対してだけでなく属性値に対しても文字列検索を行うことができる。

【００８７】

なお、データベース構築処理において、要素実体や属性値から固定長の２文字連鎖で部分文字列の切り出しを行うと説明したが、他の切り出し方法、例えば特開平８－２４９３５４号公報「文書検索装置および単語索引作成方法および文書検索方法」に記載の方法等でも構わない。

【００８８】

また、データベース検索処理において、検索条件式をＸＰａｔｈ式で与えるとして説明したが、同様の意味を持つ他のクエリ言語であっても本発明を適用することは可能である。

【００８９】

このような構成とすることによって、本実施の形態では、構造化文書の登録の際に、構造化文書に含まれる文書構造を示す要素名と祖先パス名と属性名の一覧と、それらの構造化文書中での出現位置情報のインデクスを生成することにより、構造化文書構造の全文検索のみならず、文書構造を示す検索式に示される文書を効率的に検索することができる。

【００９０】

なお、本実施の形態では、構造化文書を登録する際に、文書構造を解析して辞書データおよび出現位置索引データを構築して構造化文書を登録する構成と、受け付けた文書構造を示す検索式に示される文書を辞書データおよび出現位置索引データに基づいて登録文書を効率的に検索する構成とを同時に実現する形態としたが、登録する機能のみの構成、あるいは検索のみする構成として実現してもよい。

【００９１】

なお、本実施の形態では、構造化文書を登録する際に、要素と祖先パスに対する辞書データならびに出現位置索引データを生成して登録する構成と、この構成に属性に対する辞書データならびに出現位置索引データを生成して登録する構成と、さらにこの構成に要素や属性値のテキストに対する出現位置索引データを生成して登録する構成とを同時に実現する形態としたが、要素と祖先パスのみを対象として登録する構成、あるいは、この構成に属性を対象に加えて登録する構成、あるいは、さらにこの構成にテキストを対象に加えて登録する構成として実現してもよい。

【００９２】

（実施の形態２）

次に、本実施の形態２におけるデータベース装置の構成および動作について説明する。本実施の形態におけるデータベース装置の構成は、図１に示した実施の形態１と同じである。ただし、祖先パス登録部１０４が、文書に出現する各祖先パス名に対してではなく、祖先パス名をいくつかに分割した各部分祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて祖先パス名辞書１０８に登録すること、出現情報登録部１０６が、各要素の出現する文書番号、文字位置、文字数、祖先パス名ＩＤ列、分岐順、空要素順の情報を、要素名ＩＤをキーにして要素出現情報格納部１１１へ、各要素の出現する文書番号、文字位置、文字数、要素名ＩＤ、分岐順、空要素順の情報を、祖先パス名ＩＤ列をキーにして祖先パス出現情報格納部１１２へ、各属性の出現する文書番号、文字位置、文字数、要素名ＩＤ、祖先パス名ＩＤ列、分岐順、空要素順の情報を、属性名ＩＤをキーにして属性出現情報格納部１１３へ、要素内のテキストから切り出した部分文字列、および要素の持つ属性の値から切り出した部分文字列に関して、出現する文書番号、文字位置、祖先パス名ＩＤ列、要素名ＩＤ、属性名ＩＤ、分岐順、空要素順の情報を、部分文字列をキーにしてテキスト出現情報格納部１１４へそれぞれ格納する、という点が実施の形態１とは異なっている。

【００９３】

はじめに、文書登録（データベース構築）処理の動作について図２を用いて説明する。なお、実施の形態１と同様の処理を行う部分については詳細な説明を省略する。

【００９４】

ステップ２２０１において、入力文書解析部１０２は構造化文書を１つ読み込みユニークな文書番号を割り振った後、ステップ２２０２で、この構造化文書の論理構造を解析す

る。その際、実施の形態１の場合の処理に加え、各要素に関する「空要素順」の情報についても求める。「空要素順」とは、同じ親要素を持つ兄弟要素のうちで、先頭の要素であるかもしくは直前の兄弟要素が空要素（子孫要素を含めて要素実体のテキストを全く持たない要素）でない要素の場合には１、それ以外の場合（すなわち、直前の兄弟要素が空要素である場合）には、直前の兄弟要素の空要素順の値に１を加えた値を、最上位階層から当該要素に至るまでの各階層において求め並べたものである。

【００９５】

図２４は、本発明の実施の形態２における空要素順の説明する図である。図２４において、文書の木構造と空要素順の一例を示している。また、斜線模様の四角い枠は要素実体のテキストを含む要素２８０１、２８０４、２８０５を、無地の四角い枠は要素実体を含まない空要素２８０２、２８０３を、各要素の右肩に記された“１／２／３”のような文字列は、各要素の空要素順２８０６の情報を表している。

【００９６】

兄弟要素２８０１～２８０４の空要素順の最初の２つの数字“１／２”は祖先要素の空要素順にあたる部分で兄弟要素に共通であり、末尾の数字ｎが各要素毎に変わりうる。要素２８０１は兄弟要素の中の先頭要素であるのでｎ＝１、要素２８０２は直前の要素２８０１が空要素ではないのでｎ＝１、要素２８０３は直前の要素２８０２が空要素なので１を加えてｎ＝２、要素２８０４は直前の要素２８０３が空要素なのでさらに１を加えてｎ＝３となる。したがって、兄弟要素２８０１～２８０４の空要素順はそれぞれ、“１／２／１”、“１／２／１”、“１／２／２”、“１／２／３”となる。なお、空要素順の表記方法はこれに限らない。例えば、１以外の値を持つ階層の深さとその値を並べる方法でもよく、そのような方法で空要素順２８０６（“１／２／３”）を表記すれば、深さ１の値は１なので省略、深さ２の値が２、深さ３の値が３、したがって“２：２，３：３”となる。空要素がほとんど現れない文書、すなわち、空要素順の値がほとんど１である文書を扱う場合には、後者の表記方法の方が出現位置索引ファイルのサイズを小さくできる。

【００９７】

次に、入力文書解析部１０２の解析結果をうけて、当該文書に出現する各要素について以下の処理を繰り返す。

【００９８】

ステップ２２０３では実施の形態１と同様の処理を行う。

【００９９】

ステップ２２０４において、祖先パス名登録部１０４は、着目要素の祖先パス名を３階層毎に分割していき、分割後の各部分祖先パス名が祖先パス名辞書１０８に登録済みかどうかを調べ、登録済みであれば対応する祖先パス名ＩＤを取得し、登録されていなければ新たに祖先パス名ＩＤ（＞０）を割り当てて祖先パス名辞書１０８に登録する。なお、祖先パス名の深さが３階層以下ならば、祖先パス名ＩＤ列は実施の形態１の場合と同じように単一の祖先パス名ＩＤとなる。図２５は、本発明の実施の形態２における祖先パス名と祖先パス名ＩＤ列の例を示す図である。図２５において、祖先パス名２９０１と対応する祖先パス名ＩＤ列２９０２、および、祖先パス名辞書１０８の内容２９０３の例を示している。このように祖先パス名を分割して各部分祖先パス名に祖先パス名ＩＤを割り当てることで、当該要素の祖先要素や他の要素の処理において登録済の祖先パス名ＩＤを共用することができる。また、祖先パス名ＩＤの異なり数を小さくでき、祖先パス名辞書１０８のサイズを小さくすることが可能となる。

【０１００】

なお、本実施例では祖先パス名を３階層毎に分割する例を示したが、分割の方法はこれに限らない。例えば４階層毎に分割したり、階層の深さによって分割幅を変化させたりするようにしても構わない。また、祖先パス名ＩＤ列の区切り文字として“：”を用いたが他の区切り文字でも構わない。

【０１０１】

もし、着目要素が属性を持っているならば、ステップ２２０５～ステップ２２０６にお

いて実施の形態１と同様の処理を行う。

【０１０２】

ステップ２２０７において、出現情報登録部１０６は、着目要素に関する要素出現情報を、要素名ＩＤをキーとして要素出現情報格納部１１１に登録する。要素出現情報は、文書番号、着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置および文字数、祖先パス名ＩＤ列、分岐順、空要素順の６種類の値の組から構成される。なお、「文字位置」は、タグを除く当該文書内の全てのテキストをつなげた文字列において先頭から何文字目にあたるかで表す。また、着目要素が要素実体のテキストを全く含まない要素（＝空要素）である場合には、着目要素以降に初めて現れる（タグ以外の）テキストの先頭文字位置を着目要素の先頭文字位置とみなす。要素出現情報の一例を図２６に示す。図２６は、本発明の実施の形態２における要素出現情報を説明する図である。実施の形態１と異なるのは、要素出現情報に単一の祖先パス名ＩＤではなく１つ以上の祖先パス名ＩＤを区切り文字で連ねた祖先パス名ＩＤ列が記録されることと、空要素順の情報が含まれることである。

【０１０３】

ステップ２２０８において、出現情報登録部１０６は、着目要素に関する祖先パス出現情報（すなわち、文書番号、着目要素（子孫要素も含む）に含まれる（タグ以外の）テキストの先頭文字位置および文字数、要素名ＩＤ、分岐順、空要素順の６種類の値の組）を、祖先パス名ＩＤ列をキーとして祖先パス出現情報格納部１１２に登録する。祖先パス出現情報の一例を図２７に示す。図２７は、本発明の実施の形態２における祖先パス出現情報を説明する図である。実施の形態１と異なるのは、祖先パス出現情報に空要素順の情報が含まれることと、単一の祖先パス名ＩＤではなく１つ以上の祖先パス名ＩＤを区切り文字で連ねた祖先パス名ＩＤ列をキーとして祖先パス名出現情報が祖先パス出現情報格納部１１２登録されることである。

【０１０４】

もし、着目要素が属性を持っているならば、ステップ２２０９～ステップ２２１０において、出現情報登録部１０６は着目要素の各属性に関する属性出現情報を、属性名ＩＤをキーとして属性出現情報格納部１１３に登録する。属性出現情報は、文書番号、属性値の先頭文字位置および文字数、祖先パス名ＩＤ列、要素名ＩＤ、分岐順、空要素順の７種類の値の組から構成される。実施の形態１と異なるのは、属性出現情報に単一の祖先パス名ＩＤではなく１つ以上の祖先パス名ＩＤを区切り文字で連ねた祖先パス名ＩＤ列が記録されることと、空要素順の情報が含まれることである。

【０１０５】

ステップ２２１１において、出現情報登録部１０６は、着目要素の実体内容のテキストから部分文字列の切り出しを行い、テキスト出現情報を、切り出された部分文字列をキーとしてテキスト出現情報格納部１１４に登録する。ただし、テキスト出現情報は属性値ではないので、属性名ＩＤには常に０を格納する。テキスト出現情報は、文書番号、切り出された部分文字列の先頭文字位置、祖先パス名ＩＤ列、要素名ＩＤ、属性名ＩＤ、分岐順、空要素順の７種類の値の組から構成される。実施の形態１と異なるのは、テキスト出現情報に単一の祖先パス名ＩＤではなく１つ以上の祖先パス名ＩＤを区切り文字で連ねた祖先パス名ＩＤ列が記録されることと、空要素順の情報が含まれることである。

【０１０６】

もし、着目要素が属性を持っているならば、ステップ２２１２～ステップ２２１３において、出現情報登録部１０６は、着目要素が持つ各属性の属性値文字列から部分文字列の切り出しを行い、テキスト出現情報格納部１１４に部分文字列をキーとして登録する。ステップ２２１１と同様、実施の形態１と異なるのは、テキスト出現情報に単一の祖先パス名ＩＤではなく１つ以上の祖先パス名ＩＤを区切り文字で連ねた祖先パス名ＩＤ列が記録されることと、空要素順の情報が含まれることである。

【０１０７】

以降ステップ２２１４～２２１５の処理を実施の形態１と同様に行い、文書登録（デー

タベース構築）処理が完了する。

【０１０８】

続いて、登録済みの文書群に対する検索処理に関して説明する。実施の形態１で説明した検索式と同様の形式を持つ検索式での検索処理については、検索条件解析部１１７において、祖先パス名から祖先パス名ＩＤを求めて内部条件に変換する処理を、祖先パス名から祖先パス名ＩＤ列を求めるように変更すればよい。すなわち、祖先パス名を３階層毎に分割し、祖先パス名辞書１０８を参照して分割後の各部分祖先パス名に対応する祖先パス名ＩＤを求め、それらの祖先パス名ＩＤを順に区切り文字で区切って並べ祖先パス名ＩＤ列を求める。祖先パス名ＩＤ列の形式は、文書登録処理の説明で図２５に示した例と同様であり、祖先パス名の深さが３階層以下の場合には単一の祖先パス名ＩＤとなる。また、これに伴い、実施の形態１では出現情報取得部１１８において祖先パス名ＩＤで照合していた各種処理を、祖先パス名ＩＤ列で照合するように変更することで、検索結果を求めることができるようになる。

【０１０９】

（検索式３２０１の場合）

図２８は、本発明の実施の形態２における検索式の例を示す図である。図２８に示すＸＰａｔｈ式は「最上位階層のＡ要素の子のＢ要素の子のＸ要素の兄弟要素で、Ｘ要素より後ろに現れるＹ要素」を表している。検索条件入力部１１６に入力された検索式３２０１は、検索条件解析部１１７で解析される。検索条件解析部１１７は、検索式３２０１を解析し、要素名辞書１０７、祖先パス名辞書１０８を参照して内部条件に変換し、出現情報取得部１１８に出力する。ただし、内部条件は、「Ｃ１かつ（Ｃ２またはＣ３）、ただし、Ｃｘ：｛祖先パス名ＩＤ＝２５かつ要素名ＩＤ＝１０｝、Ｃｙ：｛祖先パス名ＩＤ＝２５かつ要素名ＩＤ＝１４｝、Ｃ１：｛ＣｘとＣｙの文書番号が同一で、かつ分岐順が末尾以外等しい｝、Ｃ２：｛ＣｘよりＣｙの方が文字位置の値が大きい｝、Ｃ３：｛ＣｘとＣｙの文字位置の値が等しく、かつＣｘよりＣｙの方が空要素順の末尾の値が大きい｝」である。ここで、祖先パス名“／Ａ／Ｂ”に対応する祖先パス名ＩＤが２５、要素名“Ｘ”に対応する要素名ＩＤが１０、要素名“Ｙ”に対応する要素名ＩＤが１４である。条件Ｃ３が必要なのは、空要素とその直後に位置する要素では文字位置が同一になるため、前後関係を判断するために空要素順の値を比較しなければならないからである。

【０１１０】

図２９は、本発明の実施の形態２における検索動作を説明する図である。出現情報取得部１１８は、出現位置索引１１０を参照し、図２９に示すように、祖先パス出現情報格納部１１２における祖先パス名ＩＤ＝２５のエントリで要素名ＩＤ＝１０であるもの（Ｃｘ）、および要素名ＩＤ＝１４であるもの（Ｃｙ）を求める。続いて、Ｃ１かつ（Ｃ２またはＣ３）を満たすようなＣｘ、Ｃｙのエントリの組３３０１、３３０２を求める。例えば、（文書番号，祖先パス名ＩＤ，要素名ＩＤ，属性名ＩＤ，分岐順，空要素順）のような形式で結果データ集合３３０３として検索結果出力部１１９に出力する。検索結果出力部１１９は、求められた結果データ集合の文書実体を取得するなどして適切な形式で検索結果を出力する。

【０１１１】

なお、ＣｘおよびＣｙのエントリを求める際に、祖先パス出現情報格納部１１２における指定祖先パス名ＩＤのエントリ数と、要素出現情報格納部１１１における指定要素名ＩＤのエントリ数を比較して少ない方を選択するようにすることも可能である。

【０１１２】

このようにして、検索式３２０１に対しては、祖先パス出現情報格納部１１２または要素出現情報格納部１１１を参照して求めた２つの要素の出現位置が同じだった場合（すなわち２つの要素が、空要素とその直後の要素の関係にあった場合）に、空要素順の情報を比較することによって、前後関係の曖昧さを排除し正しい検索結果を求めることができるようになる。

【０１１３】

以上説明したように、祖先パス名登録部１０４が祖先パス名を分割し、分割後の各部分祖先パス名に対してユニークな祖先パス名ＩＤを割り当てて祖先パス名辞書１０８に登録することで、祖先パス名辞書のサイズを小さくすることが可能となる。また、出現情報登録部１０６が要素出現情報格納部１１１、祖先パス出現情報格納部１１２、属性出現情報格納部１１３、テキスト出現情報格納部１１４に空要素順の情報も格納することにより、空要素とその直後の要素の開始文字位置が同じになることによる前後関係の曖昧さを排除し、正しい検索結果を求めることができる。

【０１１４】

このような構成とすることによって、本実施の形態では、構造文書の要素にテキストが全く含まれない空要素である場合には、着目要素以降に初めて現れるテキストの先頭文字位置を着目要素の先頭文字位置とみなすものである。さらに空要素の出現順を出現位置インデクスとして生成することより、構造化文書に空要素が含まれる場合だけでなく空要素が連続して含まれる場合であっても、構造化文書構造の全文検索のみならず、空要素を含む文書構造を示す検索式に示される文書を効率的に検索することができる。また、本実施の形態におけるデータベース装置は、祖先パス名を一定の条件で分割した部分パス名に基づいて祖先パス列として登録することにより、部分パスを重複して蓄積することなく、結果的に祖先パス辞書のサイズを小さくでき、また、構造化対象を多く含む構造化文書であっても、文書構造を示す検索式に示される文書を効率的に検索することができる。

【０１１５】

なお、本実施の形態では、構造化文書を登録する際に、文書構造を解析して辞書データおよび出現位置索引データを構築して構造化文書を登録する構成と、受け付けた文書構造を示す検索式に示される文書を辞書データおよび出現位置索引データに基づいて登録文書を効率的に検索する構成とを同時に実現する形態としたが、構造化文書を登録する機能のみの構成、あるいは検索のみする構成として実現してもよい。

【０１１６】

なお、本実施の形態では、構造化文書を登録する際に、テキスト要素を持たない空要素に対応する出現位置索引データを生成して登録する構成と、祖先パス名をいくつかに分割した各部分祖先パス名に対する辞書データならびに出現位置索引データを生成して登録する構成とを同時に実現する形態としたが、空要素のみを対象として登録する構成、あるいは、祖先パス名のみを対象として登録する構成として実現してもよい。

【０１１７】

（実施の形態３）

次に、本実施の形態３におけるデータベース装置の構成および動作について説明する。図３０は、本発明の実施の形態３におけるデータベース装置の構成を示すブロック図である。図３０において、、要素出現情報格納部１１１、祖先パス出現情報格納部１１２、属性出現情報格納部１１３、テキスト出現情報格納部１１４に格納されている情報のグループ化を行う出現情報グループ化部３４０１が追加されている点が、実施の形態１および実施の形態２の構成とは異なる。

【０１１８】

はじめに、文書登録（データベース構築）処理の動作について説明する。図３１は、本発明の実施の形態３におけるデータベース装置の文書登録処理の手順を示す流れ図である。図３１において、ステップ２２０１～２２１５までの処理は実施の形態２の場合と同じであるので、説明を省略する。

【０１１９】

最後のステップ３５０１において、出現情報グループ化部３４０１は要素出現情報格納部１１１に同じ要素名ＩＤをキーとして登録されているエントリ群の中で、文書番号と文字位置を除いた４種類の情報項目（文字数、祖先パス名ＩＤ、分岐順、空要素順）の値が全て共通しているようなエントリ同士を集め、それらのエントリの数が閾値（例えば１０エントリ）を超えていたらそれらのエントリをグループ化する。次に、残ったエントリ群について、文書番号と文字位置を除いた４種類の情報項目（文字数、祖先パス名ＩＤ、分

岐順、空要素順）のうち、いずれか３種類の情報項目の値が共通しているエントリ群を求め、エントリの数が閾値を超えていたらグループ化する。なお、あるエントリが複数のグループに属する可能性があるが、その場合にはエントリ数の最も多いグループに入れるものとする。同様にしていずれか２種類の情報項目の値が共通するエントリのグループ、いずれか１種類の情報項目の値が共通するエントリのグループを順に作成し、残ったエントリは共通情報項目無しのグループとして登録する。

【０１２０】

図３２は、本発明の実施の形態３におけるグループ化された要素出現情報を説明する図である。図３２において、グループ化された要素出現情報の例を示している。グループ情報３６０１～３６０４には、各グループに属するエントリに共通する情報項目の値が格納され、個々のエントリ３６０５～３６０８には、共通しない情報項目の値のみが格納されている。第１のグループ情報３６０１は、当該グループに属する要素出現情報のエントリはどれも（文字数＝１０，祖先パス名ＩＤ＝１００，分岐順＝“１／１／１”，空要素順＝“１／１／１”）という値を共通に持つということを表している。そして、当該グループに属する個々のエントリ３６０５にはそれぞれの文書番号と文字位置だけが格納されている。第２のグループ情報３６０２は、当該グループに属する要素出現情報のエントリはどれも（祖先パス名ＩＤ＝２００，分岐順＝“１／２／１”，空要素順＝“１／２／３”）という値を共通に持ち、“＊”となっている文字数の情報項目は共通な値ではないということを表している。そして、個々のエントリ３６０６に文書番号、文字位置とともに文字数が格納されている。同様に第３のグループ情報３６０３は、当該グループに属する要素出現情報のエントリはどれも（文字数＝８，祖先パス名ＩＤ＝１５０，空要素順＝“１／２”）という値を共通に持ち、“＊”となっている分岐順の情報項目は共通な値ではないということを表している。そして、個々のエントリ３６０７に文書番号、文字位置とともに分岐順が格納されている。最後のグループ情報３６０４は共通する情報項目がないグループで、各エントリ３６０８に全ての情報項目が格納されている。

【０１２１】

祖先パス出現情報格納部１１２、属性出現情報格納部１１３、テキスト出現情報格納部１１４に格納されている各情報についても同様にして、文書番号と文字位置以外に共通な値の情報項目を持つエントリ同士のグループ化を行い、文書登録（データベース構築）処理が完了する。

【０１２２】

登録済みの文書群に対する検索処理に関しては、グループ化された各エントリの内容とグループ情報から全ての情報項目の値を復元できるので、実施の形態１や実施の形態２と同様に検索結果を求めることができる。

【０１２３】

このようにして、出現情報グループ化部３４０１を設け、出現位置索引１１０に格納されるエントリ群をグループ化し、グループ内で共通する情報項目の値を括りだし、個々のエントリには格納しないようにすることにより、索引サイズを減らすことが可能となる。

【０１２４】

このような構成とすることによって、本実施の形態では、各要素、祖先パスなどの出現位置情報についてある条件下で情報項目の値が共通する部分をグループ化、共通化してない部分とは異なる構造で格納することによって、共通する部分を重複して蓄積することなく、結果的に索引のサイズを小さくできる。

【産業上の利用可能性】

【０１２５】

本発明に係るデータベース装置は、構造化文書を効率良く検索することが可能な構造の検索用データを構築し、効率良く検索可能なデータベース装置等に適している。

【図面の簡単な説明】

【０１２６】

【図１】本発明の実施の形態１におけるデータベース装置の構成を示すブロック図

【図２】本発明の実施の形態１における文書登録処理の手順を示す流れ図
【図３】本発明の実施の形態１における登録検索対象となる構造化文書の一例を示す図
【図４】本発明の実施の形態１における構造化文書の論理構造を解析した結果である木構造の一例を示す図
【図５】本発明の実施の形態１における祖先パス名を説明する図
【図６】本発明の実施の形態１における要素名辞書の内容の一例を示す図
【図７】本発明の実施の形態１における祖先パス名辞書の内容の一例を示す図
【図８】本発明の実施の形態１における属性名辞書の内容の一例を示す図
【図９】本発明の実施の形態１における文字位置を説明する図
【図１０】本発明の実施の形態１における要素出現情報を説明する図
【図１１】本発明の実施の形態１における祖先パス出現情報を説明する図
【図１２】本発明の実施の形態１における属性出現情報を説明する図
【図１３】本発明の実施の形態１におけるテキスト出現情報を説明する図
【図１４】本発明の実施の形態１における検索式の例を示す図
【図１５】本発明の実施の形態１におけるデータベース装置の検索処理の手順を示す流れ図
【図１６】本発明の実施の形態１におけるデータベース装置の検索動作を説明する図
【図１７】本発明の実施の形態１におけるデータベース装置の検索動作を説明する図
【図１８】本発明の実施の形態１におけるデータベース装置の検索動作を説明する図
【図１９】本発明の実施の形態１におけるデータベース装置の検索動作を説明する図
【図２０】本発明の実施の形態１におけるデータベース装置の検索動作を説明する図
【図２１】本発明の実施の形態１におけるデータベース装置の検索動作を説明する図
【図２２】本発明の実施の形態１におけるデータベース装置の検索動作を説明する図
【図２３】本発明の実施の形態１におけるデータベース装置の検索動作を説明する図
【図２４】本発明の実施の形態２における空要素順の説明に用いる図
【図２５】本発明の実施の形態２における祖先パス名と祖先パス名ＩＤ列の例を示す図
【図２６】本発明の実施の形態２における要素出現情報を説明する図
【図２７】本発明の実施の形態２における祖先パス出現情報を説明する図
【図２８】本発明の実施の形態２における検索式の例を示す図
【図２９】本発明の実施の形態２における検索動作を説明する図
【図３０】本発明の実施の形態３におけるデータベース装置の構成を示すブロック図
【図３１】本発明の実施の形態３におけるデータベース装置の文書登録処理の手順を示す流れ図
【図３２】本発明の実施の形態３におけるグループ化された要素出現情報を説明する図
【図３３】従来の構造化文書管理装置の構成図
【図３４】従来の構造化文書管理装置における要素管理テーブルの例を示す図
【図３５】従来の構造化文書管理装置における文字列索引の例の一部を示す図
【図３６】従来の構造化文書管理装置における検索処理を説明する図

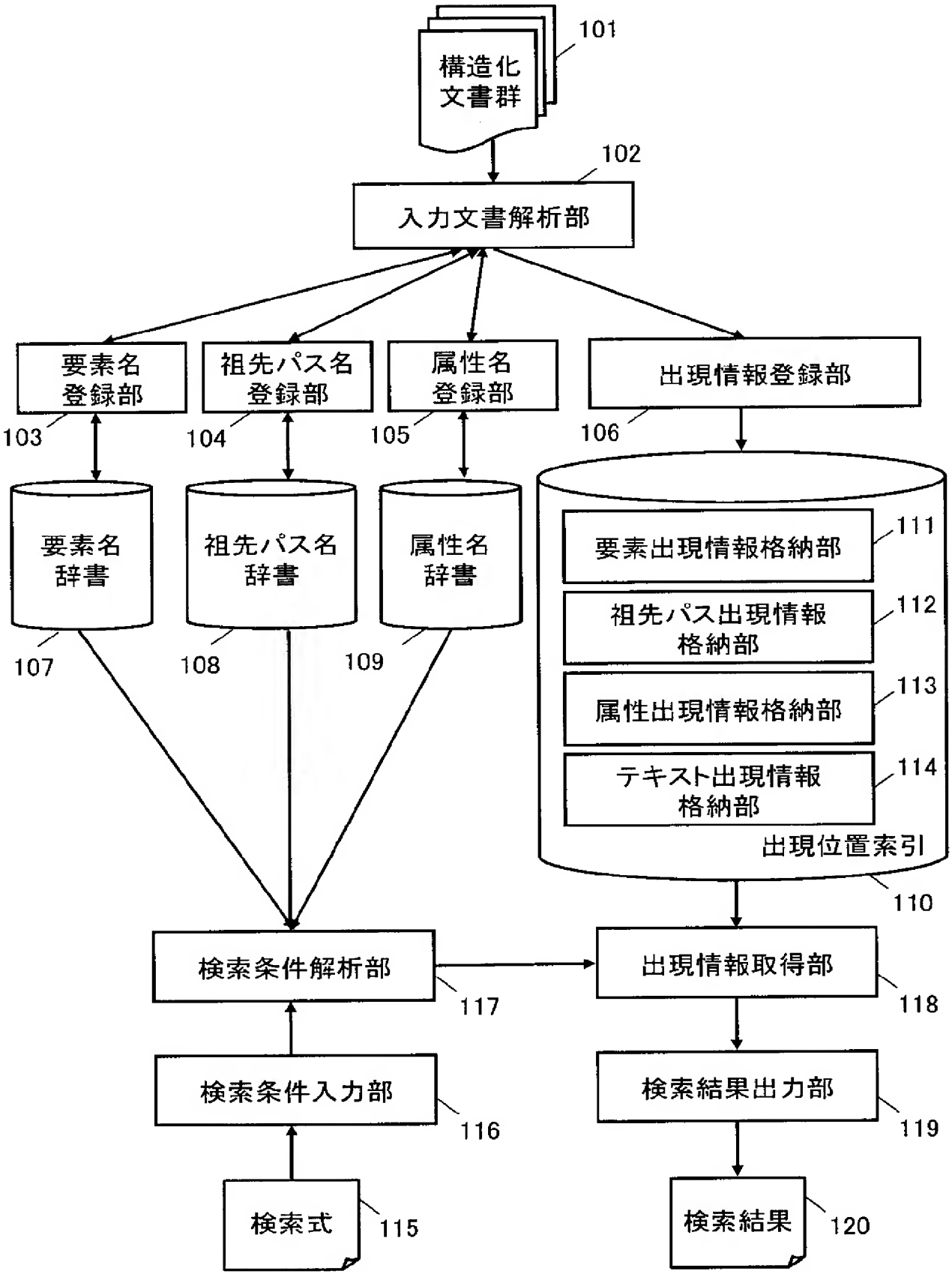
【符号の説明】

【０１２７】

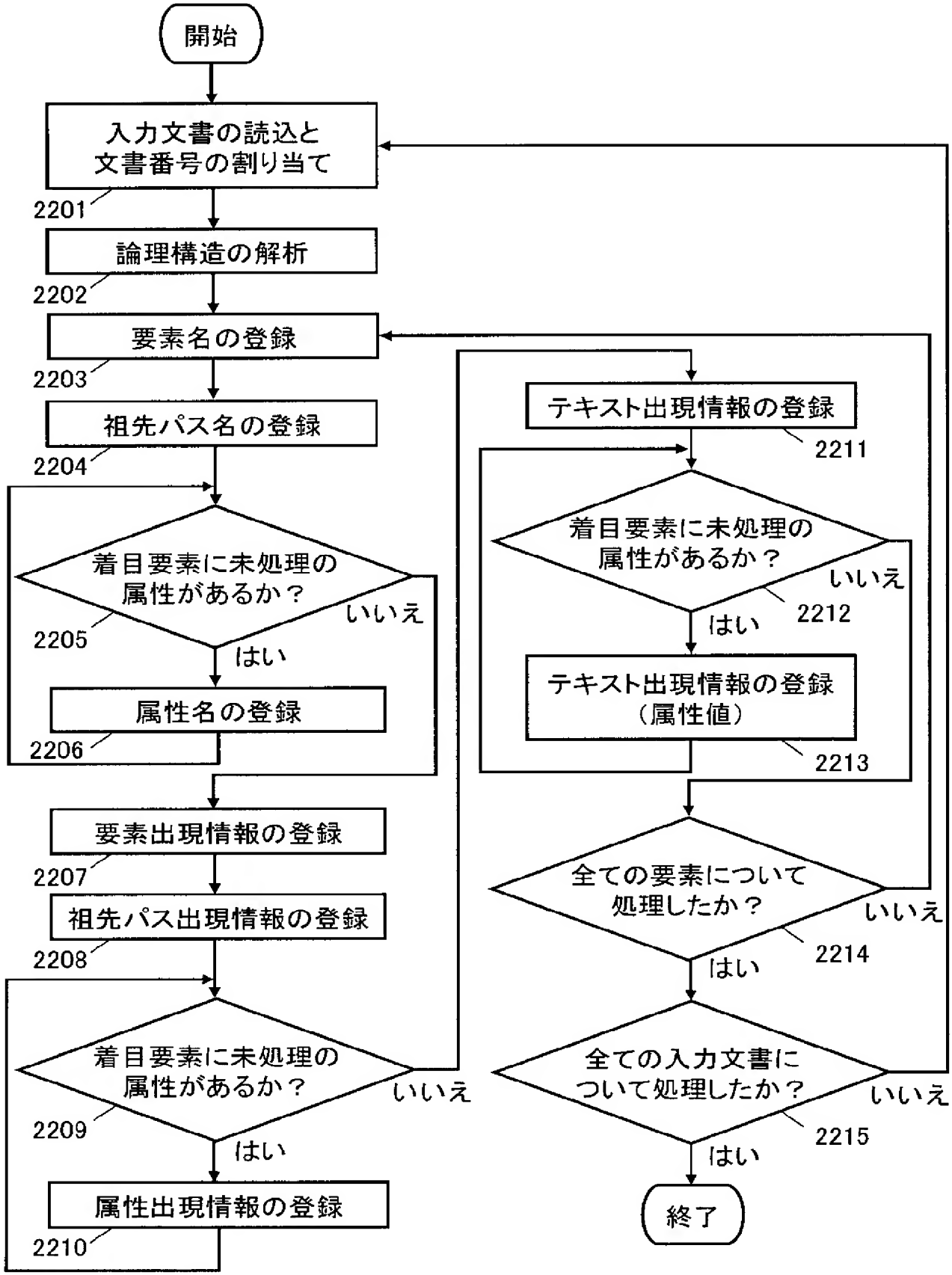
１０１　構造化文書群
１０２　入力文書解析部
１０３　要素名登録部
１０４　祖先パス名登録部
１０５　属性名登録部
１０６　出現情報登録部
１０７　要素名辞書

１０８　　祖先パス名辞書
１０９　　属性名辞書
１１０　　出現位置索引
１１１　　要素出現情報格納部
１１２　　祖先パス出現情報格納部
１１３　　属性出現情報格納部
１１４　　テキスト出現情報格納部
１１５　　検索式
１１６　　検索条件入力部
１１７　　検索条件解析部
１１８　　出現情報取得部
１１９　　検索結果出力部
１２０　　検索結果
３４０１　　出現情報グループ化部

【書類名】図面
【図１】


構造化
文書群
101
102
入力文書解析部
要素名
登録部
103
祖先パス名
登録部
104
属性名
登録部
105
出現情報登録部
106
要素名
辞書
107
祖先パス名
辞書
108
属性名
辞書
109
要素出現情報格納部
111
祖先パス出現情報
格納部
112
属性出現情報格納部
113
テキスト出現情報
格納部
114
出現位置索引
110
検索条件解析部
117
出現情報取得部
118
検索条件入力部
116
検索結果出力部
119
検索式
115
検索結果
120

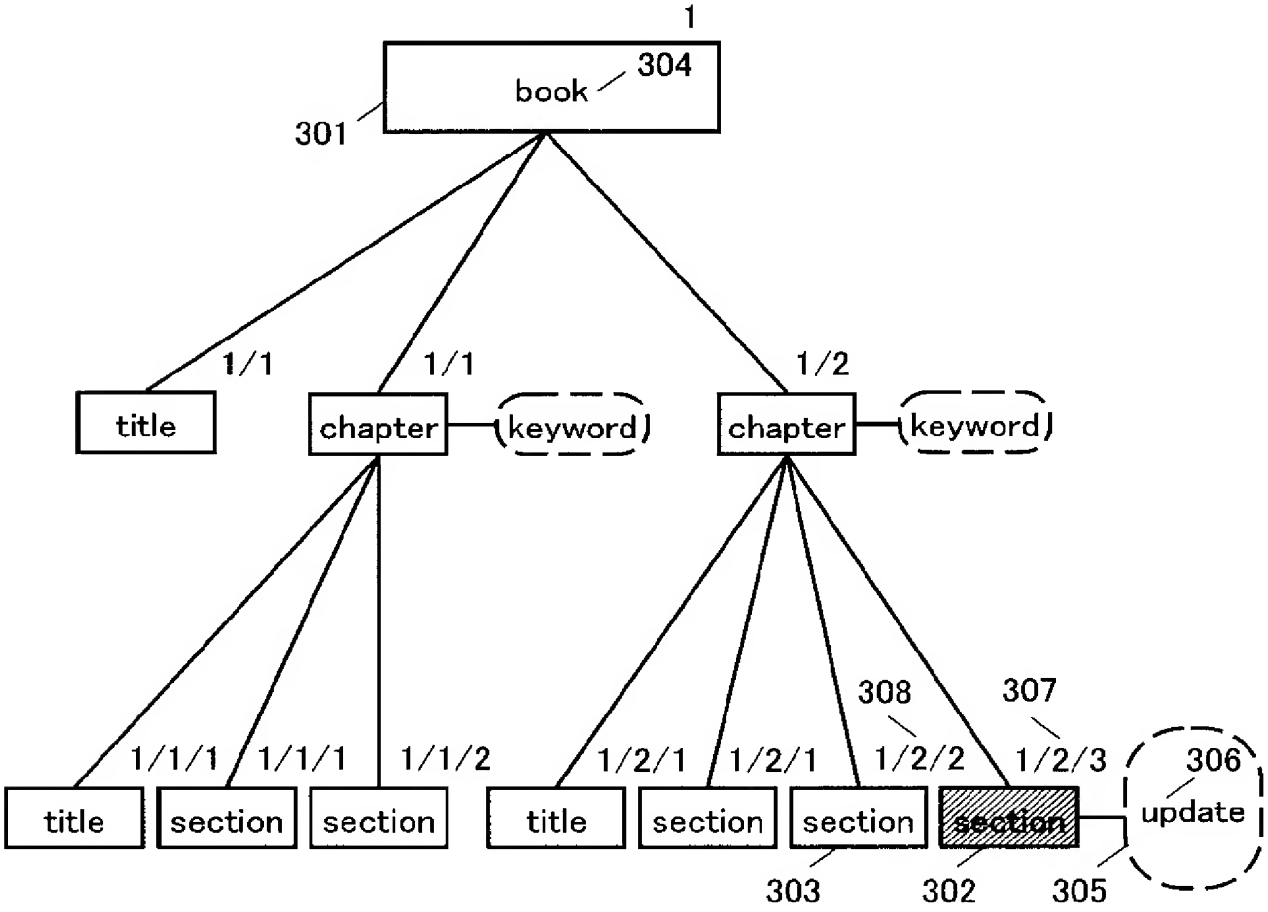
【図２】


開始
入力文書の読込と
文書番号の割り当て
2201
論理構造の解析
2202
要素名の登録
2203
祖先パス名の登録
2204
着目要素に未処理の
属性があるか？
2205
いいえ
はい
属性名の登録
2206
要素出現情報の登録
2207
祖先パス出現情報の登録
2208
着目要素に未処理の
属性があるか？
2209
いいえ
はい
属性出現情報の登録
2210
テキスト出現情報の登録
2211
着目要素に未処理の
属性があるか？
2212
いいえ
はい
テキスト出現情報の登録
（属性値）
2213
全ての要素について
処理したか？
2214
いいえ
はい
全ての入力文書に
ついて処理したか？
2215
いいえ
はい
終了

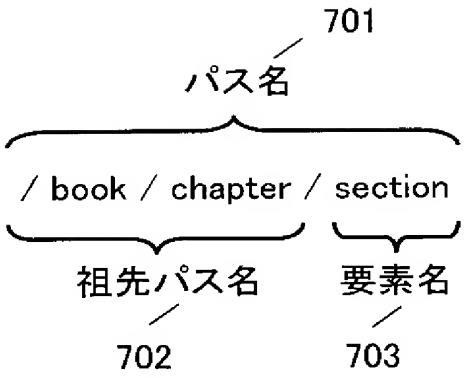
【図３】

```
<book>
  <title>文書検索</title>
  <chapter keyword="歴史">
    <title>文書検索の歴史</title>
    <section>キーワード検索は、・・・</section>
    <section>その後、全文検索が・・・</section>
  </chapter>
  <chapter keyword="索引">
    <title>索引方式</title>
    <section>最長一致切り出しによる・・・</section>
    <section>n－gram索引方式は・・・</section>
    <section update="200406">新たに極大単語索引方式が・・・</section>
  </chapter>
</book>
```

【図４】



【図５】



【図６】

要素名ID	要素名
1	book
2	title
3	chapter
4	section

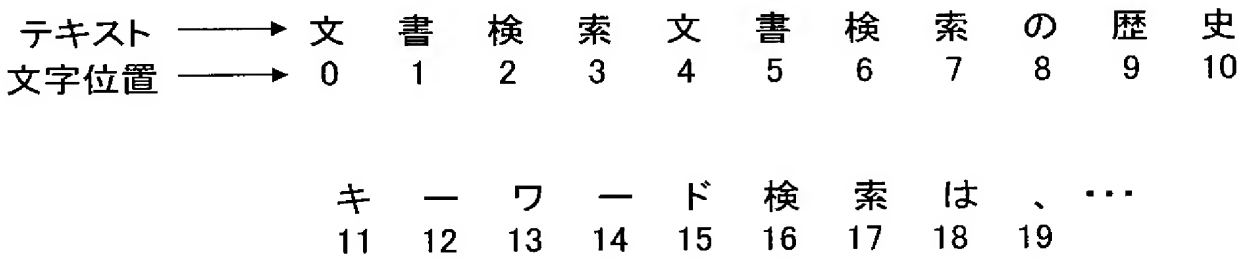
【図７】

祖先パス名ID	祖先パス名
1	/
2	/book
3	/book/chapter

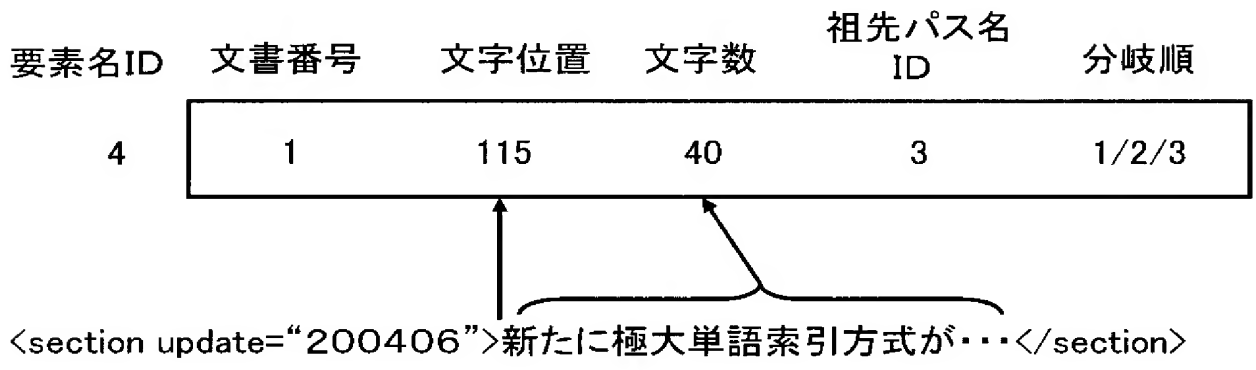
【図８】

属性名ID	属性名
1	keyword
2	update

【図９】



【図１０】



【図１１】

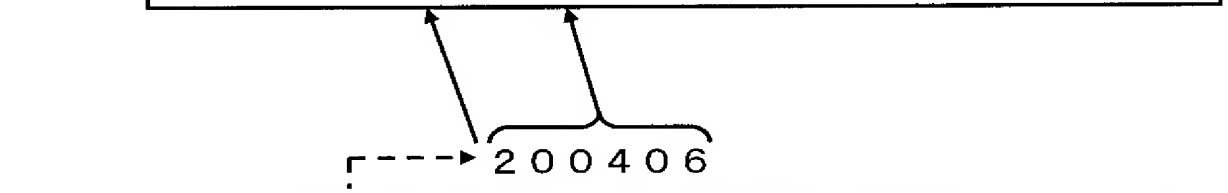
祖先パス出現情報

祖先パス名ID	文書番号	文字位置	文字数	要素名ID	分岐順
3	1	115	40	4	1/2/3

【図１２】

属性出現情報

属性名ID	文書番号	文字位置	文字数	祖先パス名ID	要素名ID	分岐順
2	1	115	6	3	4	1/2/3



<section update="200406">新たに極大単語索引方式が・・・</section>

【図１３】

テキスト出現情報

部分文字列	文書番号	文字位置	祖先パス名ID	要素名ID	属性名ID	分岐順
“新た”	1	115	3	4	0	1/2/3
“たに”	1	116	3	4	0	1/2/3
“に極”	1	117	3	4	0	1/2/3
“極大”	1	118	3	4	0	1/2/3
“大単”	1	119	3	4	0	1/2/3
“単語”	1	120	3	4	0	1/2/3

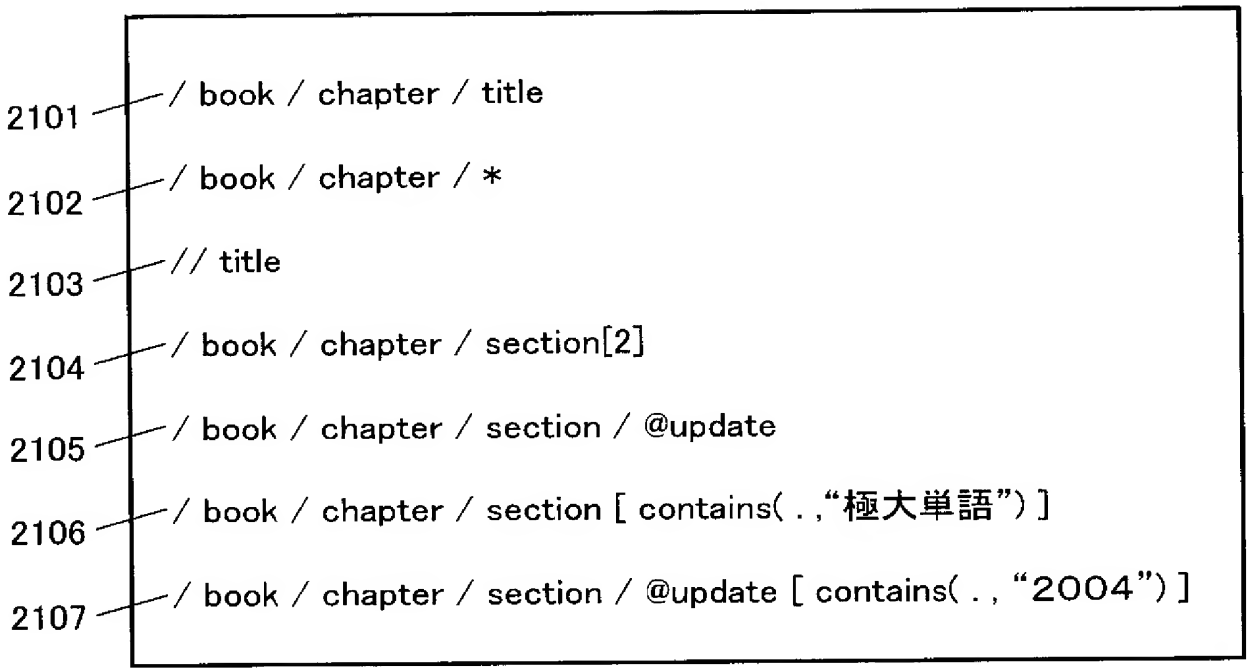
1201

属性値でない場合は0
属性値の場合は属性名ID（≠0）

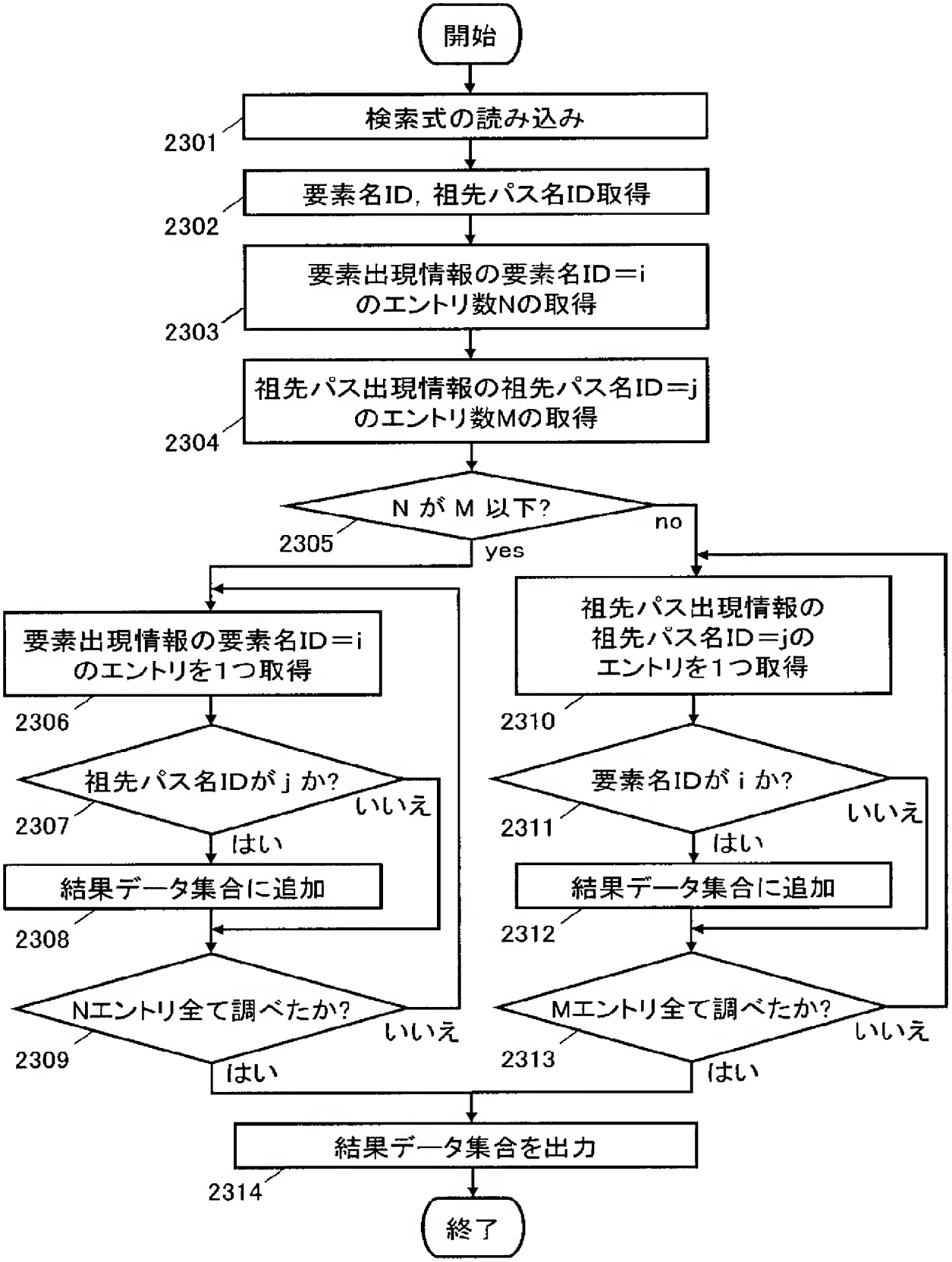
1202

“20”	1	115	3	4	2	1/2/3
“00”	1	116	3	4	2	1/2/3
“04”	1	117	3	4	2	1/2/3
“40”	1	118	3	4	2	1/2/3
“06”	1	119	3	4	2	1/2/3

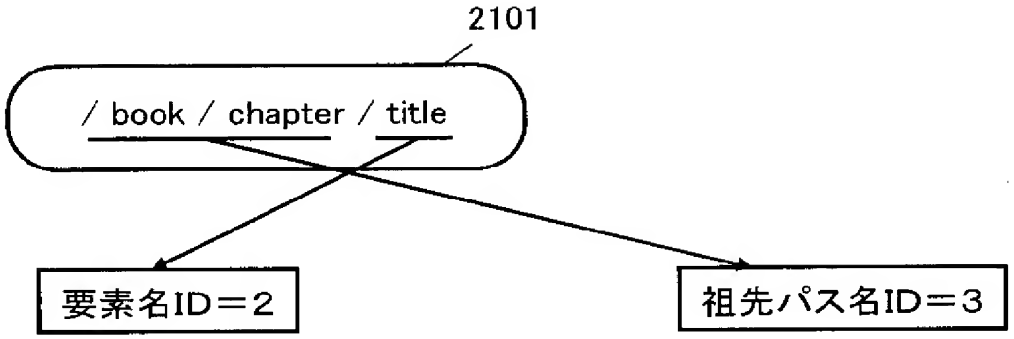
【図１４】


2101
/ book / chapter / title
2102
/ book / chapter / *
2103
// title
2104
/ book / chapter / section[2]
2105
/ book / chapter / section / @update
2106
/ book / chapter / section [contains(. ,“極大単語”)]
2107
/ book / chapter / section / @update [contains(. , “2004”)]

【図１５】


開始
検索式の読み込み
2301
要素名ID，祖先パス名ID取得
2302
要素出現情報の要素名ID＝i
のエントリ数Nの取得
2303
祖先パス出現情報の祖先パス名ID＝j
のエントリ数Mの取得
2304
N が M 以下?
2305
no
yes
要素出現情報の要素名ID＝i
のエントリを1つ取得
2306
祖先パス名IDが j か?
2307
いいえ
はい
結果データ集合に追加
2308
Nエントリ全て調べたか?
2309
いいえ
はい
祖先パス出現情報の
祖先パス名ID＝jの
エントリを1つ取得
2310
要素名IDが i か?
2311
いいえ
はい
結果データ集合に追加
2312
Mエントリ全て調べたか?
2313
いいえ
はい
結果データ集合を出力
2314
終了

【図１６】



要素出現情報

要素名ID	文書番号	文字位置	文字数	祖先パス名ID	分岐順
	：				
2	1	0	4	2	1/1
	1	4	7	3	1/1/1
	1	59	4	3	1/2/1
	2	0	6	2	1/1
	4	0	8	2	1/1
	4	24	6	3	1/1/1
	7	0	5	2	1/1
	9	7	4	3	1/1/1
3	：				

1301

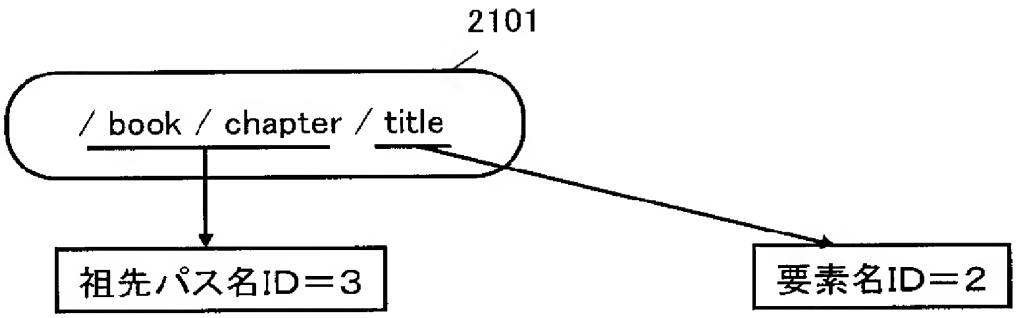
8 エントリ

結果データ集合

1302

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{（1，3，2，0，1/1/1），
（1，3，2，0，1/2/1），
（4，3，2，0，1/1/1），
（9，3，2，0，1/1/1）}

【図１７】



祖先パス出現情報

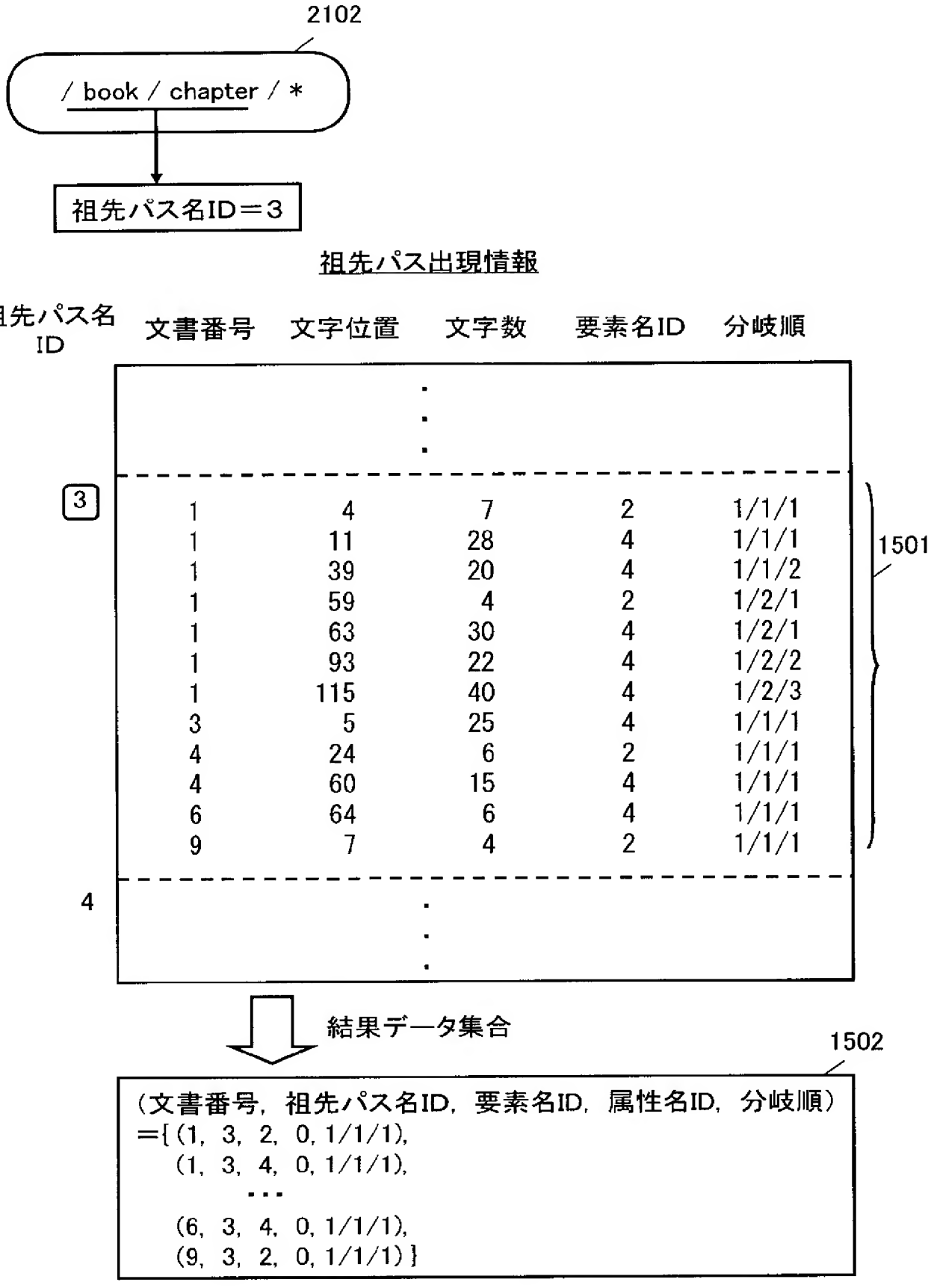
祖先パス名ID	文書番号	文字位置	文字数	要素名ID	分岐順	
	：					
3	1	4	7	2	1/1/1	1401
	1	11	28	4	1/1/1	
	1	39	20	4	1/1/2	
	1	59	4	2	1/2/1	
	1	63	30	4	1/2/1	
	1	93	22	4	1/2/2	
	1	115	40	4	1/2/3	12エントリ
	3	5	25	4	1/1/1	
	4	24	6	2	1/1/1	
	4	60	15	4	1/1/1	
	6	64	6	4	1/1/1	
	9	7	4	2	1/1/1	
4	：					

結果データ集合

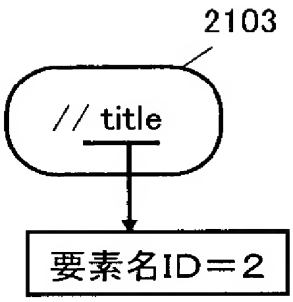
1402

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{（1，3，2，0，1/1/1），
（1，3，2，0，1/2/1），
（4，3，2，0，1/1/1），
（9，3，2，0，1/1/1）}

【図１８】


2102
/ book / chapter / *
祖先パス名ID＝3
祖先パス出現情報
祖先パス名ID
文書番号
文字位置
文字数
要素名ID
分岐順
3
1 4 7 2 1/1/1
1 11 28 4 1/1/1
1 39 20 4 1/1/2
1 59 4 2 1/2/1
1 63 30 4 1/2/1
1 93 22 4 1/2/2
1 115 40 4 1/2/3
3 5 25 4 1/1/1
4 24 6 2 1/1/1
4 60 15 4 1/1/1
6 64 6 4 1/1/1
9 7 4 2 1/1/1
1501
4
結果データ集合
1502
（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
={(1, 3, 2, 0, 1/1/1),
(1, 3, 4, 0, 1/1/1),
...
(6, 3, 4, 0, 1/1/1),
(9, 3, 2, 0, 1/1/1)}

【図１９】



要素出現情報

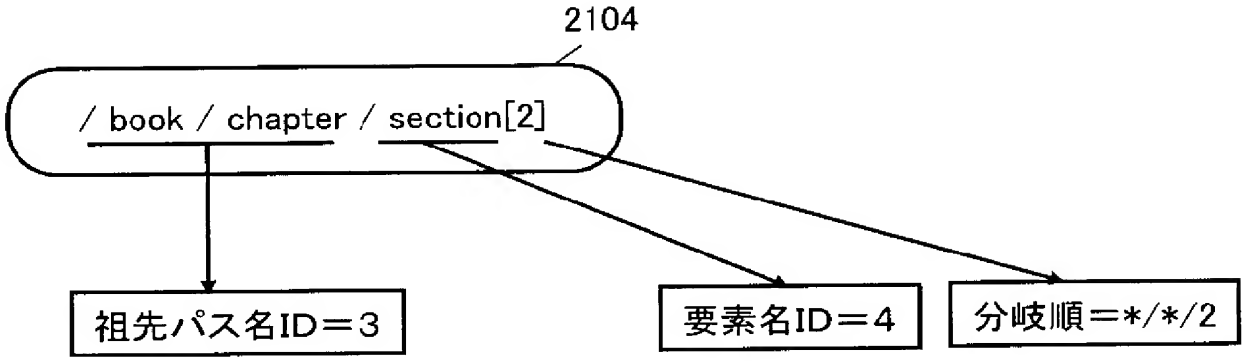
	要素名ID	文書番号	文字位置	文字数	祖先パス名ID	分岐順	
		・					
		・					
		・					
2	1	0	4	2	1/1		1601
	1	4	7	3	1/1/1		
	1	59	4	3	1/2/1		
	2	0	6	2	1/1		
	4	0	8	2	1/1		
	4	24	6	3	1/1/1		
	7	0	5	2	1/1		
	9	7	4	3	1/1/1		
3		・					
		・					
		・					

結果データ集合

1602

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{（1，2，2，0，1/1　），
（1，3，2，0，1/1/1），
・・・
（7，2，2，0，1/1　），
（9，3，2，0，1/1/1）}

【図２０】



祖先パス出現情報

祖先パス名ID	文書番号	文字位置	文字数	要素名ID	分岐順
	1	・・・			
3	1	4	7	2	1/1/1
	1	11	28	4	1/1/1
	1	39	20	4	1/1/2
	1	59	4	2	1/2/1
	1	63	30	4	1/2/1
	1	93	22	4	1/2/2
	1	115	40	4	1/2/3
	3	5	25	4	1/1/1
	4	24	6	2	1/1/1
	4	60	15	4	1/1/1
	6	64	6	4	1/1/1
	9	7	4	2	1/1/1
4		・・・			

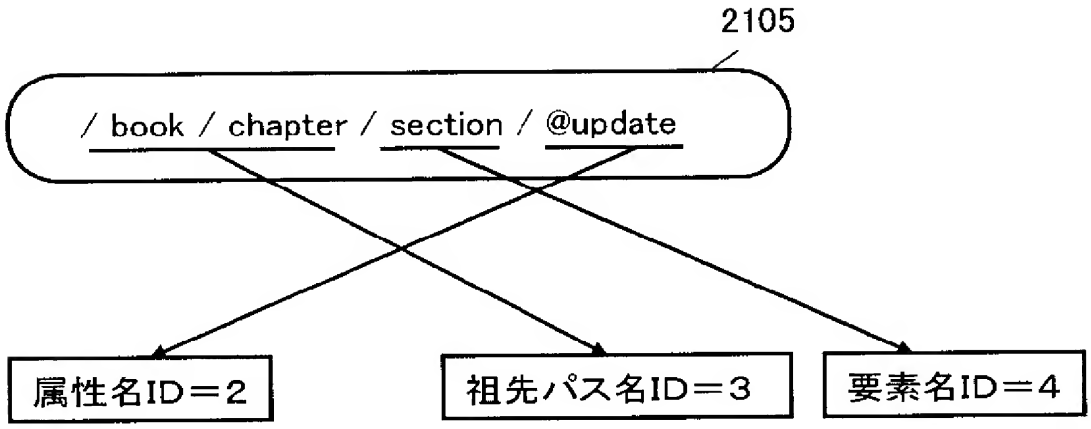
1701

結果データ集合

1702

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{（1，3，4，0，1/1/2），
（1，3，4，0，1/2/2）}

【図２１】



属性出現情報

属性名ID	文書番号	文字位置	文字数	祖先パス名ID	要素名ID	分岐順
	・					
	・					
	・					
2	1	115	6	3	4	1/2/3
	2	8	4	2	2	1/1
	5	60	6	3	4	1/1/2
	8	32	8	3	2	1/2/1
3	・					
	・					
	・					

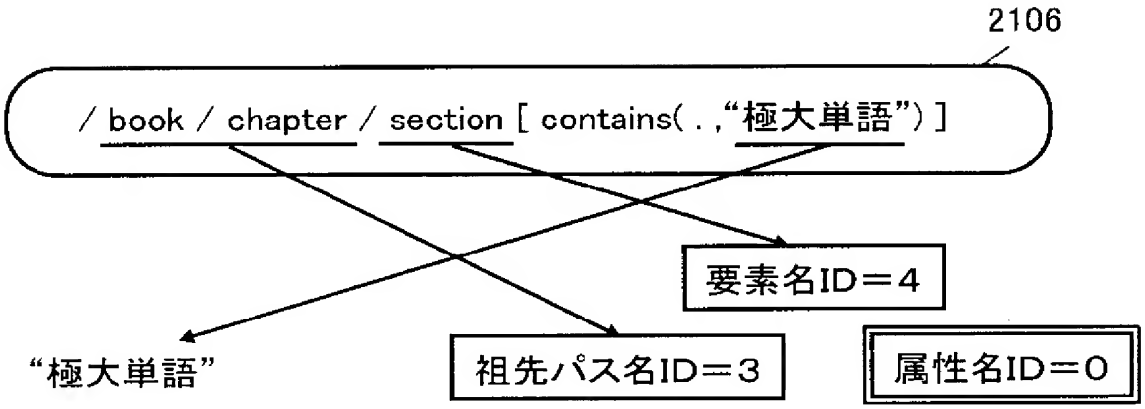
1801

結果データ集合

1802

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{（1，3，4，2，1/2/3），
（5，3，4，2，1/1/2）}

【図２２】



テキスト出現情報

部分文字列	文書番号	文字位置	祖先パス名ID	要素名ID	属性名ID	分岐順
“極大”						
	1	118	3	4	0	1/2/3
	2	86	3	4	0	1/1/1
	3	24	2	2	0	1/1
	4	62	3	4	0	1/1/1
	8	77	3	4	2	1/1/1
			・			
			・			
			・			

1901

文書番号が同じで文字位置が連接、分岐順も同じであること

“単語”						
	1	120	3	4	0	1/2/3
	3	26	2	2	0	1/1
	4	64	3	4	0	1/1/1
	8	79	3	4	1	1/1/1
			・			
			・			
			・			

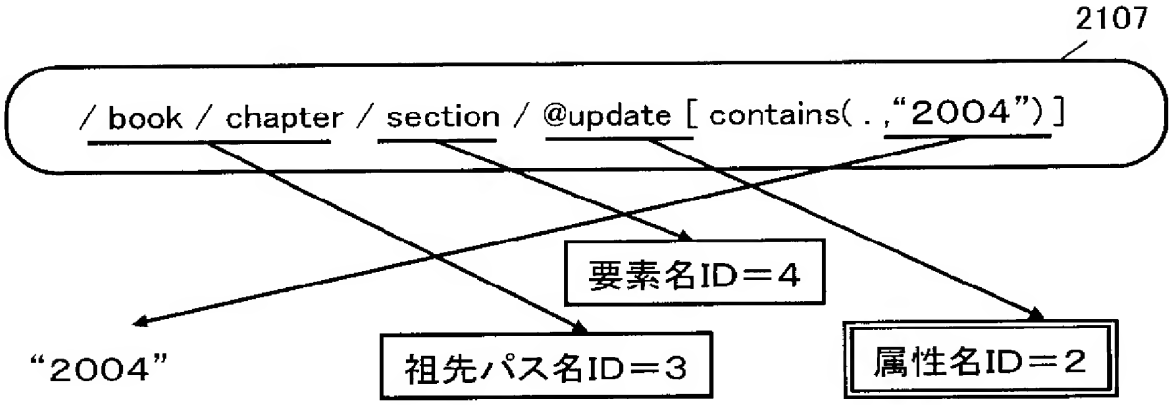
1902

結果データ集合

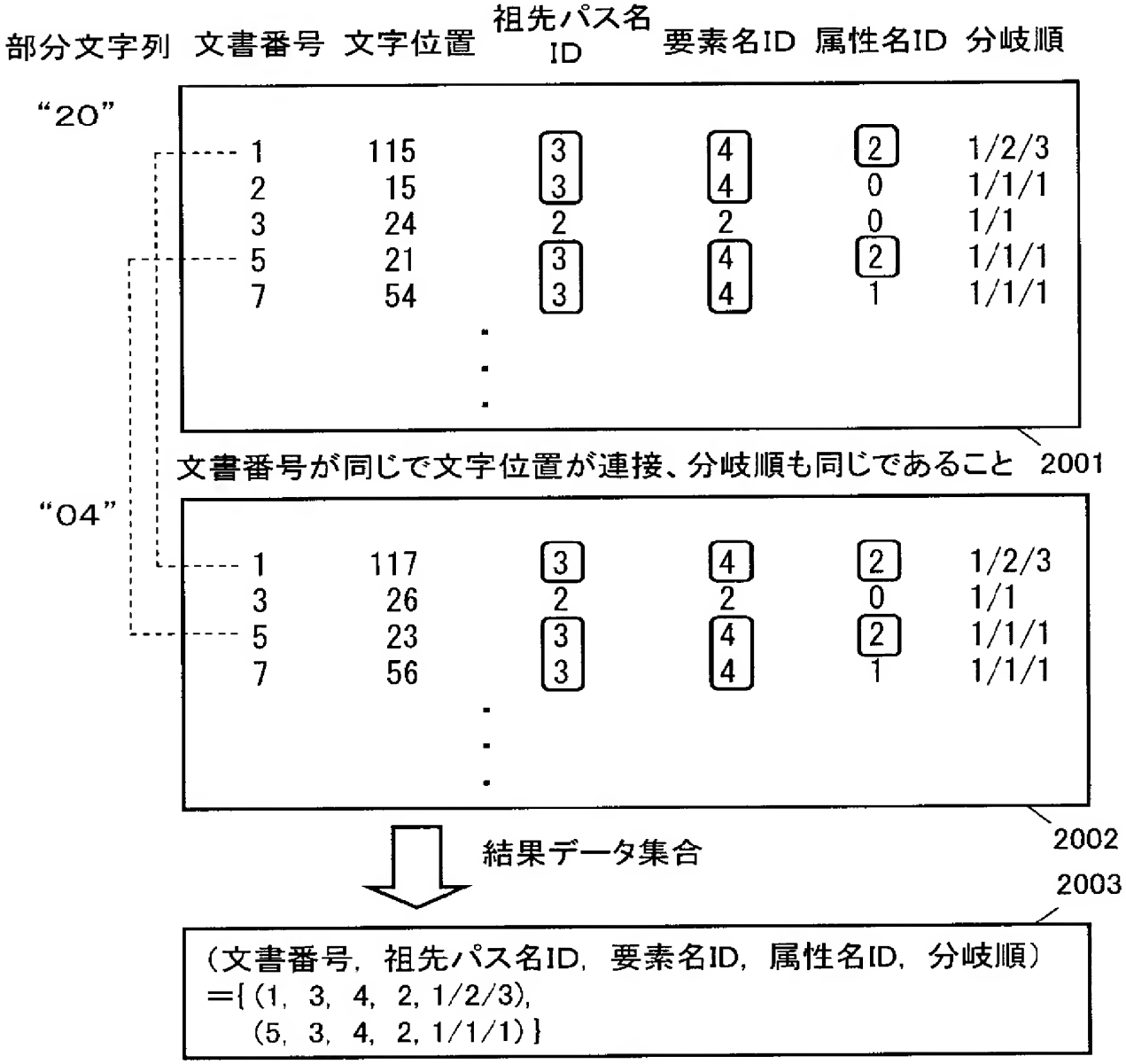
1903

（文書番号，祖先パス名ID，要素名ID，属性名ID，分岐順）
＝{（1，3，4，0，1/2/3），
（4，3，4，0，1/1/1）}

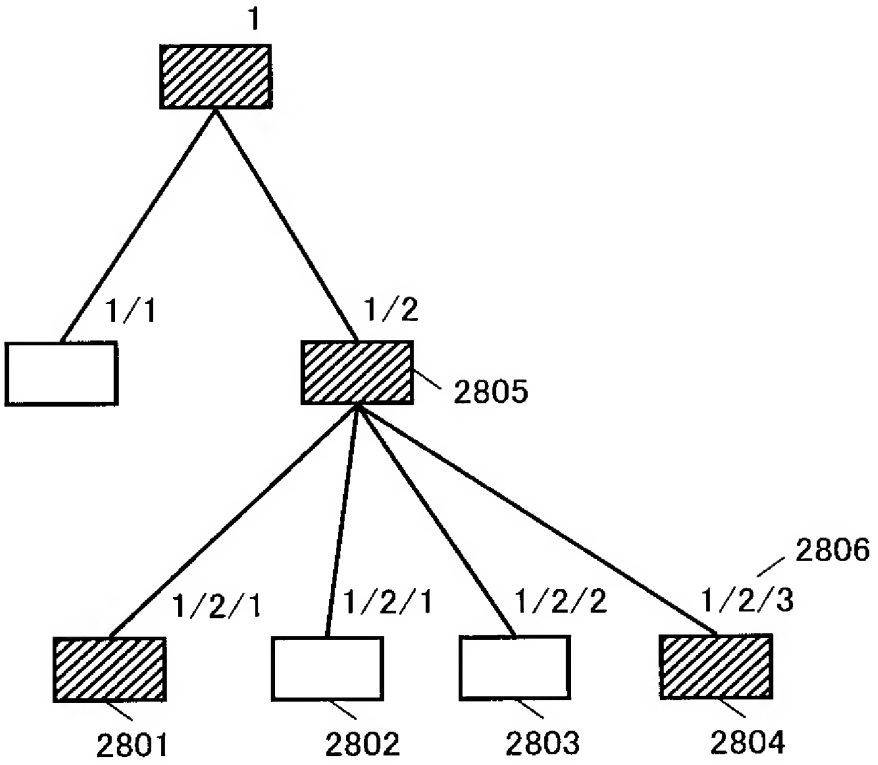
【図２３】



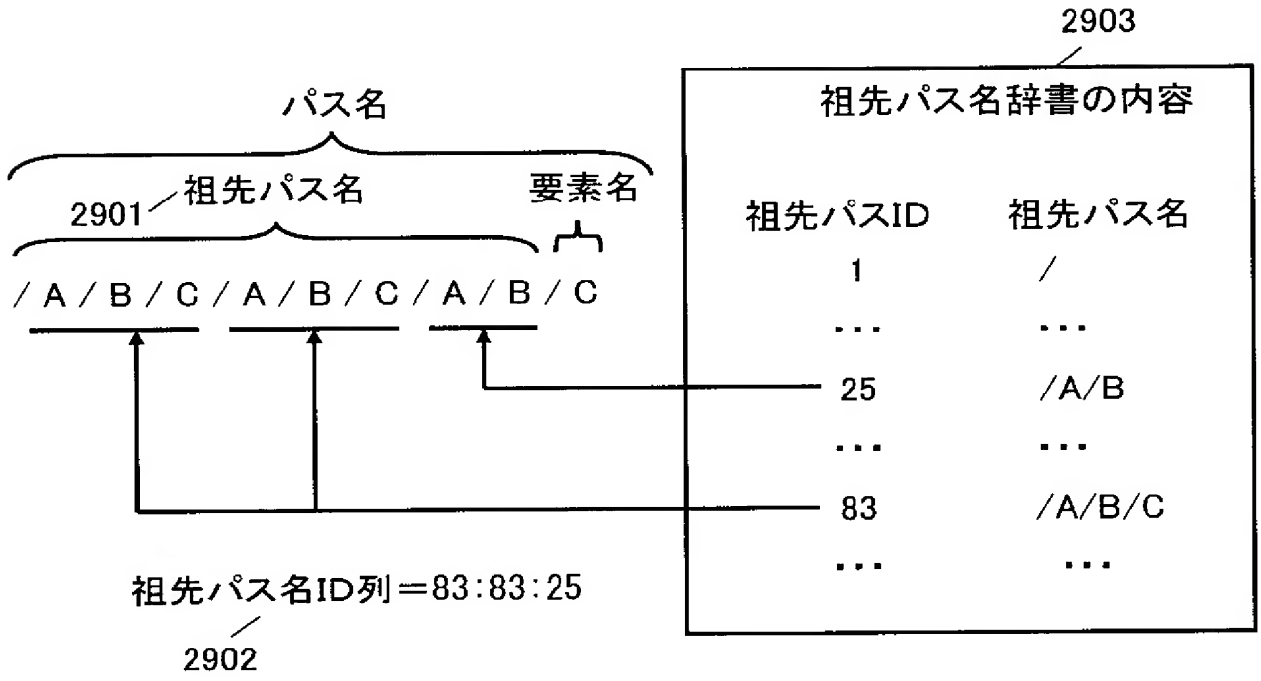
テキスト出現情報



【図２４】


1
1/1
1/2
2805
2806
1/2/1
1/2/1
1/2/2
1/2/3
2801
2802
2803
2804

【図２５】


2903
祖先パス名辞書の内容
パス名
祖先パス名
要素名
2901
/ A / B / C / A / B / C / A / B / C
祖先パスID
祖先パス名
1
/
25
/A/B
83
/A/B/C
祖先パス名ID列＝83:83:25
2902

【図２６】

要素出現情報

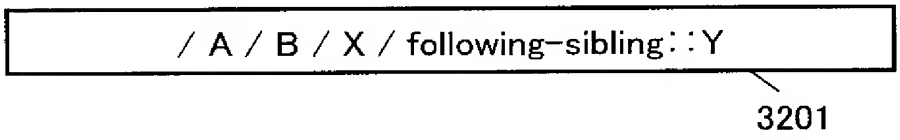
要素名ID	文書番号	文字位置	文字数	祖先パス名ID	分岐順	空要素順
10	1	100	20	83:25	1/2/3/1/1/2	1/1/2/1/2/1

【図２７】

祖先パス出現情報

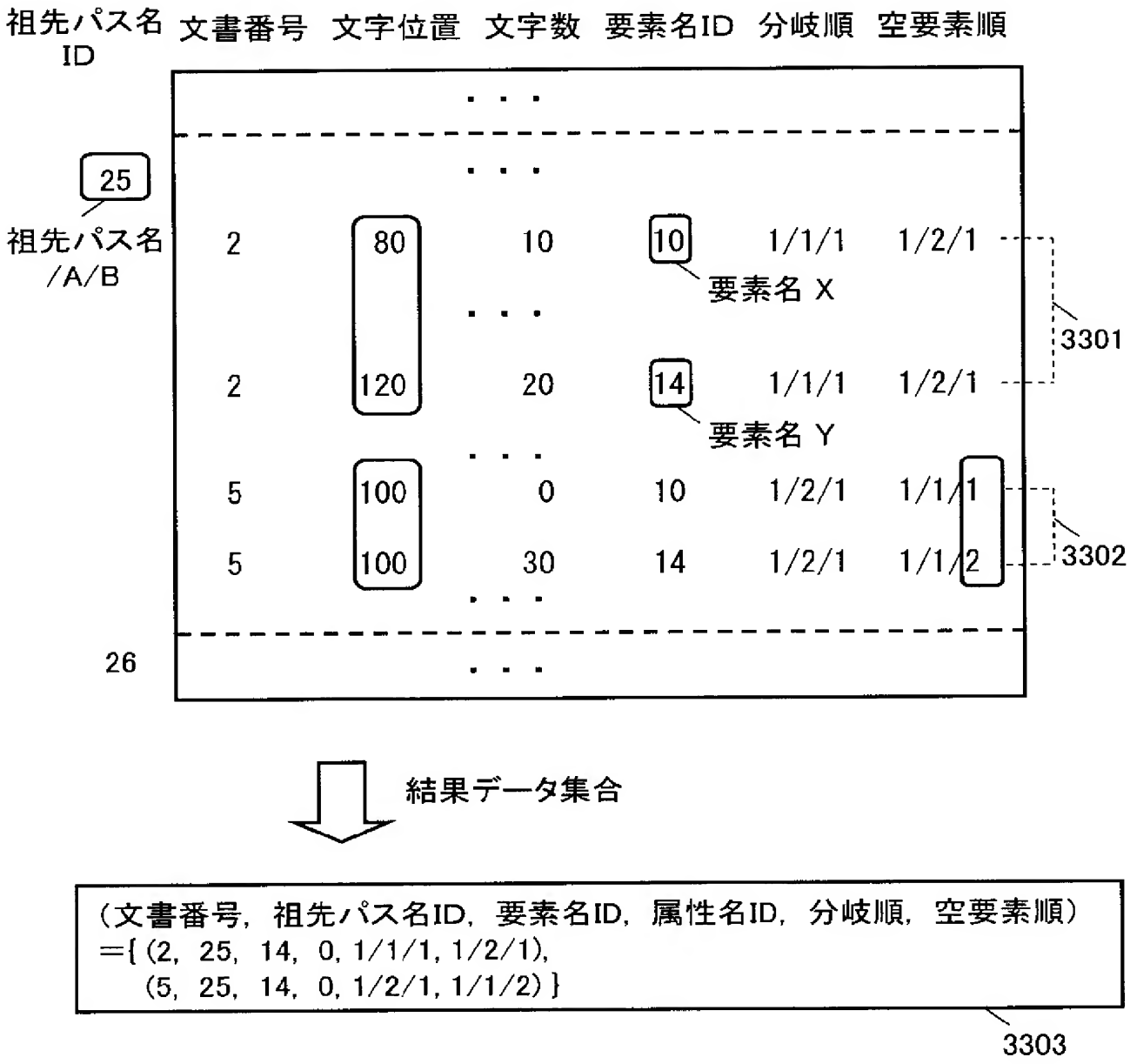
祖先パス名ID	文書番号	文字位置	文字数	要素名ID	分岐順	空要素順
83:25	1	100	20	10	1/2/3/1/1/2	1/1/2/1/2/1

【図２８】

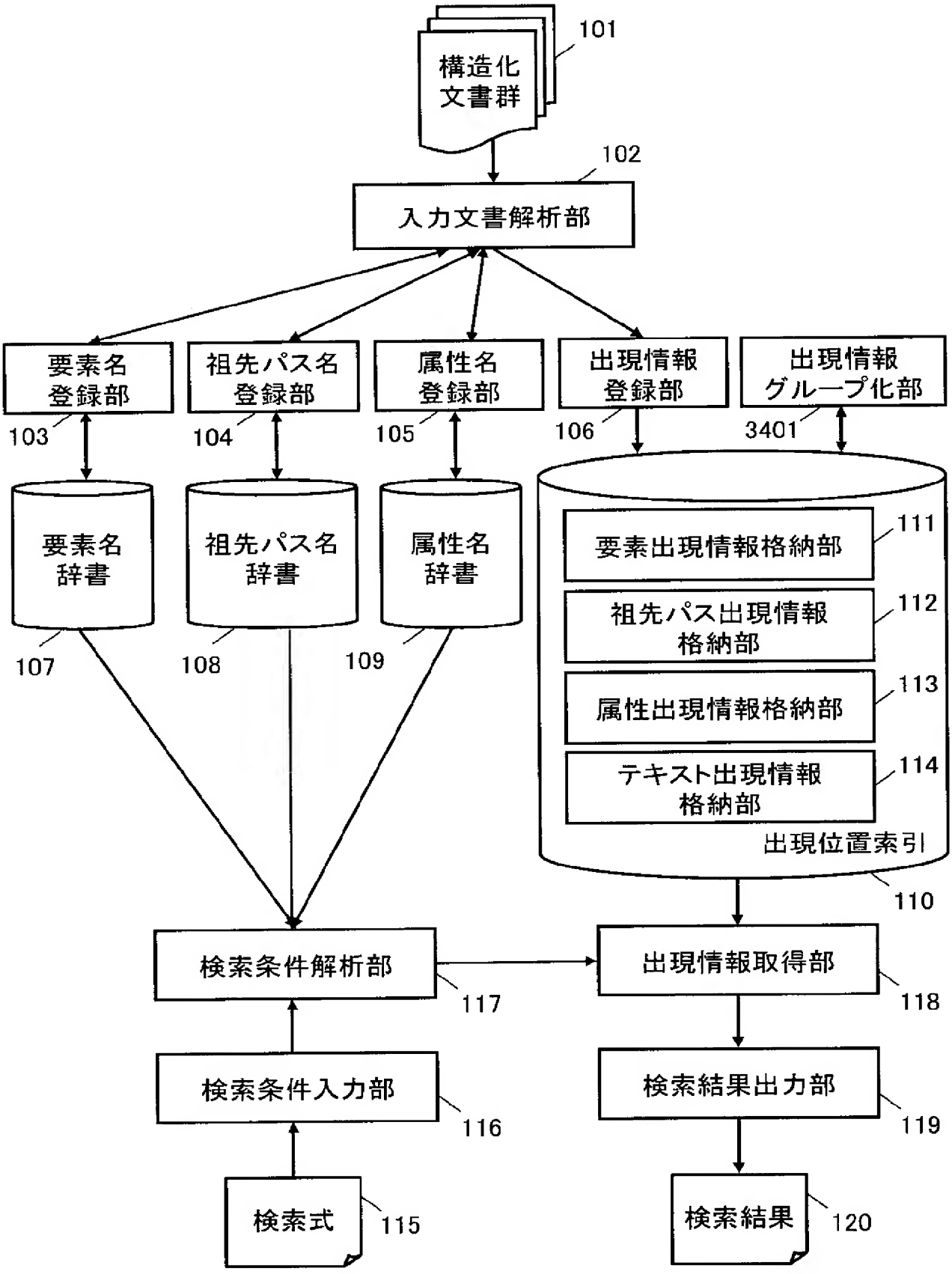


【図２９】

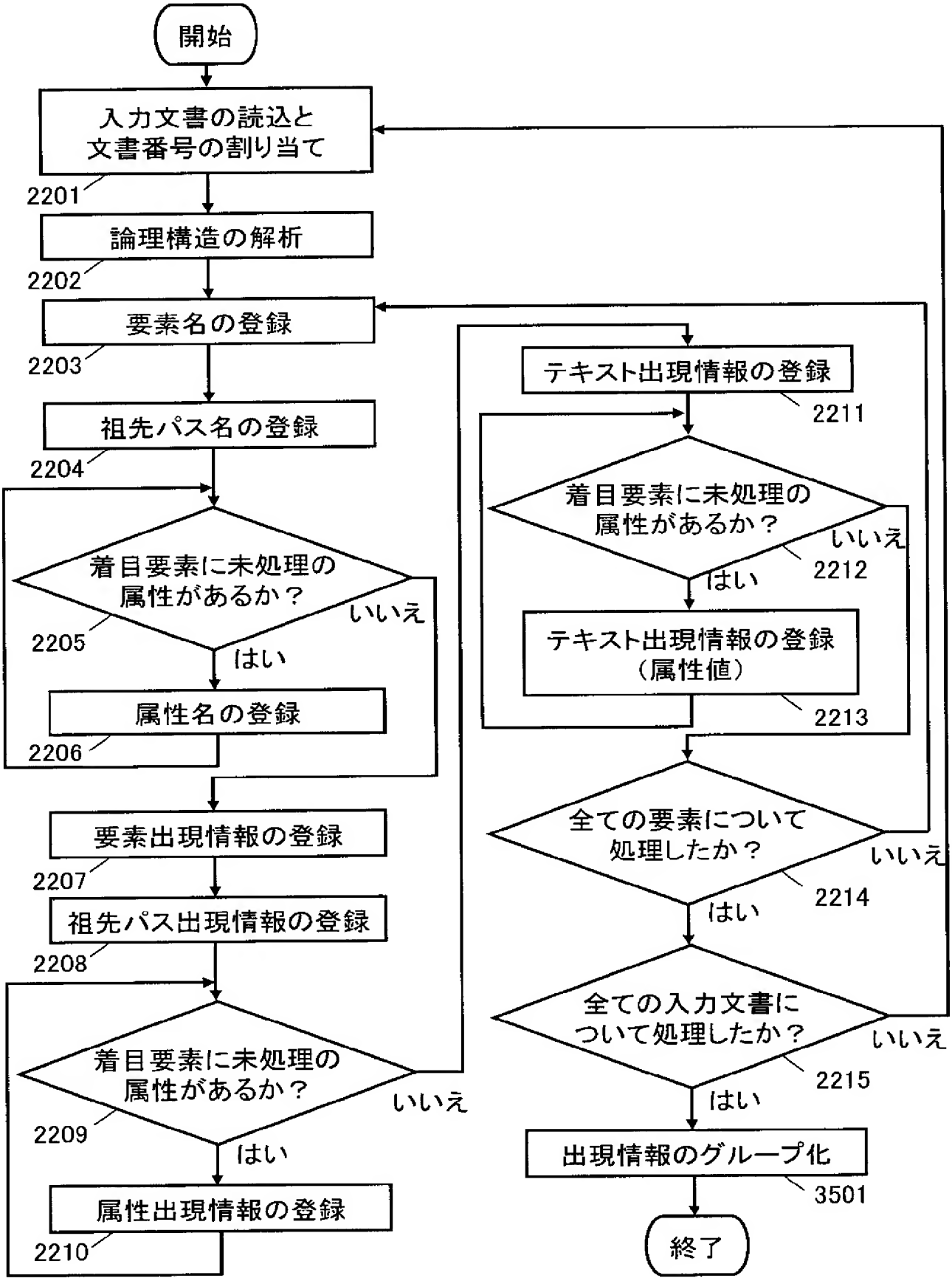
祖先パス出現情報



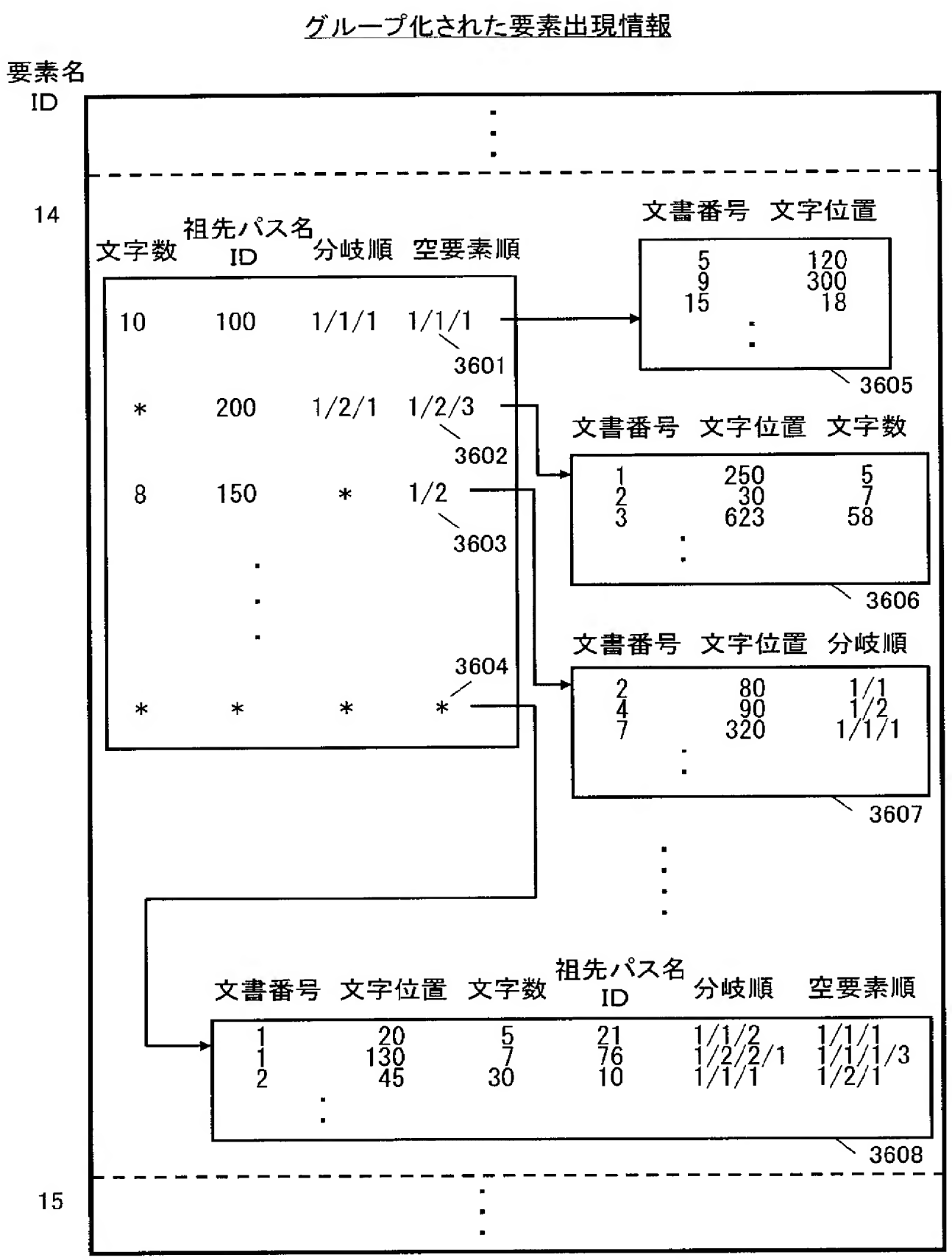
【図３０】


構造化
文書群
101
102
入力文書解析部
要素名
登録部
103
祖先パス名
登録部
104
属性名
登録部
105
出現情報
登録部
106
出現情報
グループ化部
3401
要素名
辞書
祖先パス名
辞書
属性名
辞書
107
108
109
要素出現情報格納部
111
祖先パス出現情報
格納部
112
属性出現情報格納部
113
テキスト出現情報
格納部
114
出現位置索引
110
検索条件解析部
117
出現情報取得部
118
検索条件入力部
116
検索結果出力部
119
検索式
115
検索結果
120

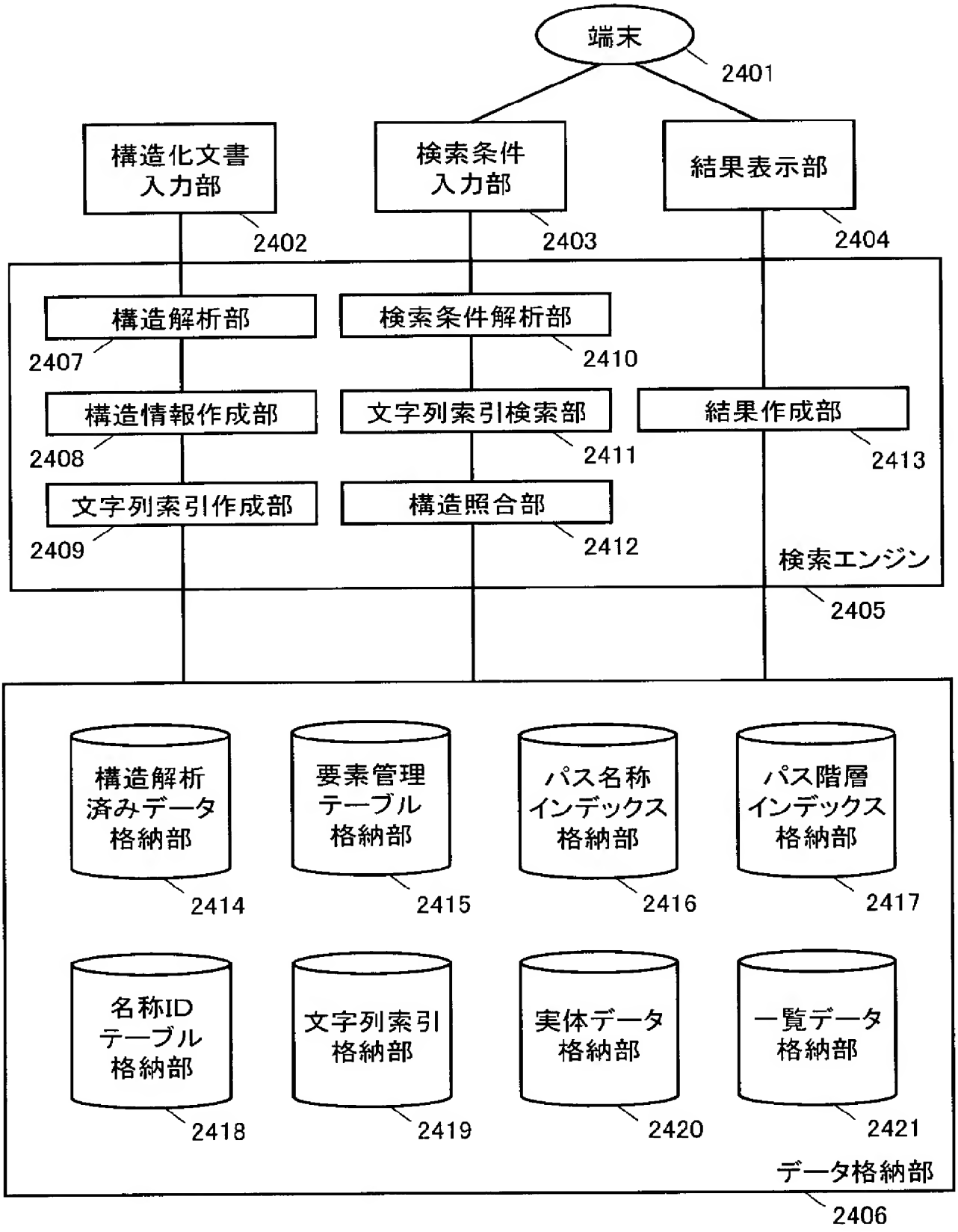
【図３１】


開始
入力文書の読込と
文書番号の割り当て
2201
論理構造の解析
2202
要素名の登録
2203
祖先パス名の登録
2204
着目要素に未処理の
属性があるか？
2205
いいえ
はい
属性名の登録
2206
要素出現情報の登録
2207
祖先パス出現情報の登録
2208
着目要素に未処理の
属性があるか？
2209
いいえ
はい
属性出現情報の登録
2210
テキスト出現情報の登録
2211
着目要素に未処理の
属性があるか？
2212
いいえ
はい
テキスト出現情報の登録
（属性値）
2213
全ての要素について
処理したか？
2214
いいえ
はい
全ての入力文書に
ついて処理したか？
2215
いいえ
はい
出現情報のグループ化
3501
終了

【図３２】


グループ化された要素出現情報
要素名
ID
14
文字数
祖先パス名
ID
分岐順
空要素順
10 100 1/1/1 1/1/1
3601
* 200 1/2/1 1/2/3
3602
8 150 * 1/2
3603
3604
* * * *
文書番号 文字位置
5 120
9 300
15 18
3605
文書番号 文字位置 文字数
1 250 5
2 30 7
3 623 58
3606
文書番号 文字位置 分岐順
2 80 1/1
4 90 1/2
7 320 1/1/1
3607
文書番号 文字位置 文字数 祖先パス名 ID 分岐順 空要素順
1 20 5 21 1/1/2 1/1/1
1 130 7 76 1/2/2/1 1/1/1/3
2 45 30 10 1/1/1 1/2/1
3608
15

【図３３】


端末
2401
構造化文書
入力部
2402
検索条件
入力部
2403
結果表示部
2404
構造解析部
2407
検索条件解析部
2410
構造情報作成部
2408
文字列索引検索部
2411
結果作成部
2413
文字列索引作成部
2409
構造照合部
2412
検索エンジン
2405
構造解析
済みデータ
格納部
2414
要素管理
テーブル
格納部
2415
パス名称
インデックス
格納部
2416
パス階層
インデックス
格納部
2417
名称ID
テーブル
格納部
2418
文字列索引
格納部
2419
実体データ
格納部
2420
一覧データ
格納部
2421
データ格納部
2406

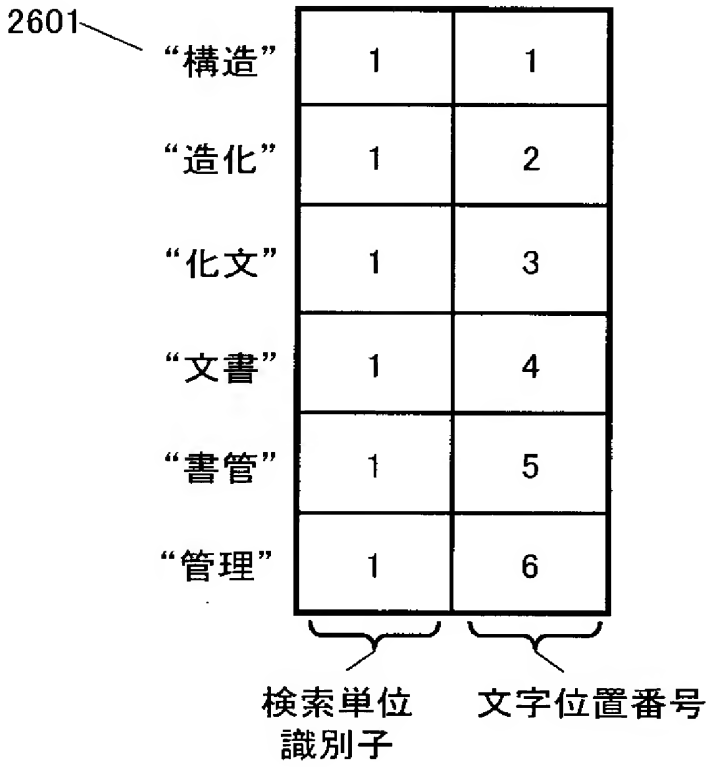
【図３４】

検索単位識別子	文書番号	パス名称ID	パス階層ID	名称ID
1	1	N2	L2	T3
2	1	N3	L2	T4
3	1	N3	L6	T4
4	1	N4	L2	T5
5	1	N7	L3	T8
6	1	N8	L3	T9
7	1	N10	L4	T11
8	1	N11	L4	T9
・ ・ ・	・・・	・・・	・・・	・・・

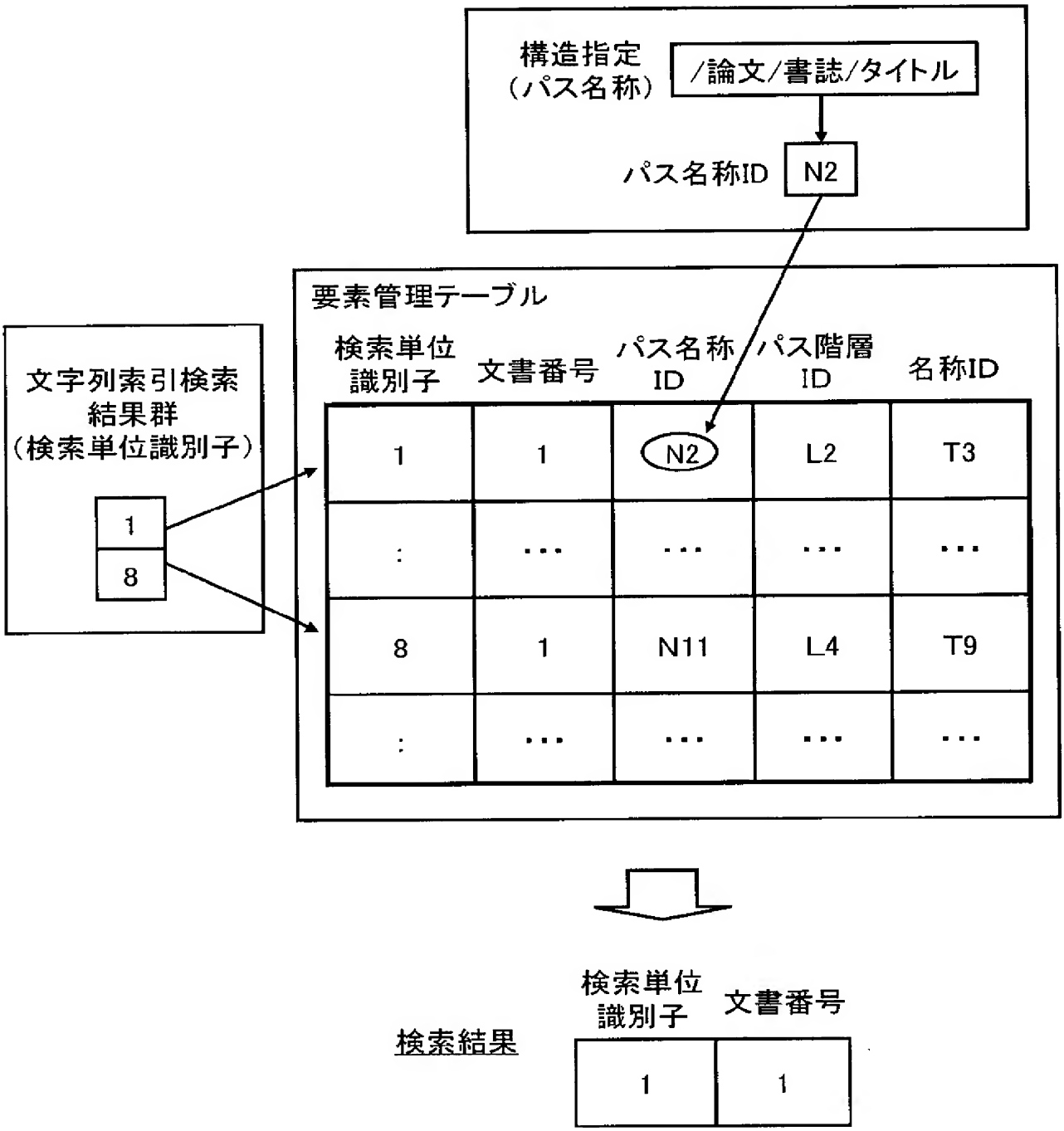
要素管理テーブル

【図３５】

<タイトル>構造化文書管理</タイトル>



【図３６】


構造指定
（パス名称）
/論文/書誌/タイトル
パス名称ID
N2
要素管理テーブル
検索単位識別子
文書番号
パス名称ID
パス階層ID
名称ID
1
1
N2
L2
T3
:
・・・
・・・
・・・
・・・
8
1
N11
L4
T9
:
・・・
・・・
・・・
・・・
文字列索引検索結果群
（検索単位識別子）
1
8
検索結果
検索単位識別子
文書番号
1
1

【書類名】要約書
【要約】
【課題】様々な検索条件での構造化文書に対する検索を効率良く行う。構造条件のみでの検索や、属性値に対する文字列検索ができるようにすることを目的とする。
【解決手段】要素の出現情報を、要素名ＩＤをキーにして格納した要素出現情報格納部と、要素の出現情報を、その要素の祖先パス名ＩＤをキーにして格納した祖先パス出現情報格納部と、属性の出現情報を、属性名ＩＤをキーにして格納した属性出現情報格納部と、要素実体のテキスト文字列、および要素の持つ属性の属性値に関する出現情報を、部分文字列をキーにして格納したテキスト出現情報格納部とを備える。
【選択図】図１

出願人履歴

０００００５８２１
19900828
新規登録

大阪府門真市大字門真１００６番地
松下電器産業株式会社